The Web: Data Integration and Cutting Edge Applications

Vasileios Lapatas

Department of Informatics

Ionian University

This dissertation is submitted for the degree of

Doctor of Philosophy

April 2016

The Web: Data Integration and Cutting Edge Applications

by

Vasileios Lapatas
Ionian University
Department of Informatics

Supervisor

Dr. Michalis Stefanidakis
Assistant Professor
Ionian University
Department of Informatics

Advisory Committee Members

Dr. Spyridon Sioutas
Associate Professor
Ionian University
Department of Informatics

Dr. Theodore Andronikos
Assistant Professor
Ionian University
Department of Informatics

Committee Members

Dr. Georgios Vouros
Professor
University of Piraeus
Department of Digital Systems

Dr. Nektarios Koziris
Professor
National Technical University of Athens
School of Electrical and Computer Engineering

Dr. Vasileios Chrisikopoulos
Professor
Ionian University
Department of Informatics

Dr. Ioannis Papadakis
Assistant Professor
Ionian University
Department of Archives, Library Science & Museology

A Thesis submitted for the degree of Doctor of Philosophy in the Department of Informatics of the School of Information & Informatics of the Ionian University

Corfu, Greece
April 2016

I would like to dedicate this thesis to my family. My father that without him none of this would be possible. My mother and my brother who never stopped believing in me.

Περίληψη

Όλα ξεκίνησαν το 1993, λιγότερο από 30 χρόνια από σήμερα, που ο παγκόσμιος ιστός γεννήθηκε στο διαδίκτυο και έγινε ελεύθερα προσβάσιμος σε όλον τον κόσμο. Σήμερα το διαδίκτυο και ο παγκόσμιος ιστός αποτελούν αναμφισβήτητα αναπόσπαστα κομμάτια της ζωής του σύγχρονου ανθρώπου. Οι τεχνολογίες διαδικτύου έχουν, κατά μεγάλο βαθμό αλλάζει την καθημερινότητα του ανθρώπου τόσο στην επαγγελματική αλλά και προσωπική του ζωή. Τα τελευταία χρόνια αυτό φαίνεται από την διάδοση των έξυπνων κινητών τηλεφώνων τα οποία προσφέρουν πρόσβαση στον παγκόσμιο ιστό από σχεδόν οποιοδήποτε σημείο στον κόσμο.

Η παρούσα διατριβή εξετάζει διάφορες τεχνολογίες διαδικτύου με έμφαση στην ολοκλήρωση δεδομένων μέσω των προηγμένων τεχνολογιών διαδικτύου. Πιο αναλυτικά, ερευνήθηκαν τα ακόλουθα ερωτήματα:

- Πως μπορούν να συμβληθούν διαδικτυακές υπηρεσίες με δεδομένα χρήστη για να δημιουργηθούν συμπεράσματα με βάση τα δεδομένα αυτά;
- Πως θα μπορούσε να επωφεληθεί ο τομέας του e-Learning από σύγχρονες τεχνολογίες διαδικτύου;
- Πως γίνεται να ενισχυθεί η αναζήτηση αρχείων σε έναν υπολογιστή με δεδομένα από το διαδίκτυο;
- Πως μπορούν να επιλυθούν πιθανά προβλήματα συγχρονισμού σε συστήματα αποθηκών δεδομένων;

- Ποια είναι η τωρινή κατάσταση σε βιολογικά δεδομένα στο διαδίκτυο και πώς θα μπορούσε να βελτιωθεί;

- Πως είναι δυνατόν να απλοποιηθεί η αναζήτηση διδακτικής ύλης για έναν τομέα όπως η βιοπληροφορική;

Από τα ερωτήματα που τέθηκαν είναι φανερό ότι η παρούσα διατριβή καλύπτει θέματα από διάφορους τομείς: ακαδημαϊκό, επιχειρησιακό, πληροφορική, βιοπληροφορική και μάθηση. Κάθε κεφάλαιο της παρούσας διατριβής αποτελεί και μία απάντηση στα ερωτήματα που τέθηκαν.

Στο κεφάλαιο 2 παρουσιάζεται μία εφαρμογή η οποία χρησιμοποιεί δεδομένα χρήστη και πιο συγκεκριμένα τους σελιδοδείκτες του φυλλομετρητή του, τροφοδοτώντας στη συνέχεια τα ανακτημένα δεδομένα σε μία υπηρεσία κατηγοριοποίησης ιστοσελίδων. Τα αποτελέσματα της υπηρεσίας αυτής αποτελούν μια αναπαράσταση των ενδιαφερόντων του χρήστη. Εφαρμογές σαν κι αυτήν θα μπορούσαν να χρησιμοποιηθούν σαν προγραμματιστικές διεπαφές για προσαρμογή ιστοσελίδων με βάση τα ενδιαφέροντα του χρήστη. Τεχνολογίες σαν κι αυτήν ανήκουν σήμερα στην καθημερινότητα των χρηστών και εταιρίες όπως η Google χρησιμοποιούν τέτοιου είδους τεχνικές με σκοπό τη βελτίωση των υπηρεσιών τους.

Το κεφάλαιο 3 αποτελεί μία ανασκόπηση προηγμένων τεχνολογιών διαδικτύου και προτείνει μία υποθετική εφαρμογή που τις αξιοποιεί με σκοπό το συνεργατικό e-Learning. Το προτεινόμενο σύστημα περιέχει στοιχεία όπως συνεργατική συγγραφή κειμένου, συνδέσμους με εξωτερικές ιστοσελίδες όπως wikis, αξιοποίηση υπηρεσιών διαδικτύου με σκοπό την αυτοματοποίηση ορισμένων λειτουργιών όπως στοχευμένη αναζήτηση και αυτοματοποιημένη σύνδεση με wikis. Επιπλέον η εφαρμογή είναι σε θέση να υποστηρίξει λειτουργία επανάληψης του μαθήματος όπου ο χρήστης θα μπορεί να παρακολουθήσει βήμα-βήμα τι έγινε στο διαδικτυακό μάθημα. Μερικές από τις προτεινόμενες λειτουργίες έχουν ήδη εφαρμοστεί στα κυρίαρχα συστήματα ηλεκτρονικής μάθησης όπως Moodle, Drupal

και Blackboard θα ήταν όμως ενδιαφέρον να υλοποιηθούν στο μέλλον τα πρόσθετα αυτοματοποιημένα στοιχεία του προτεινόμενου συστήματος.

Στο κεφάλαιο 4 παρουσιάζεται η μελέτη και η ανάπτυξη εφαρμογής που αξιοποιεί αποτελέσματα αναζήτησης σε έναν υπολογιστή με σκοπό την πρόταση παρόμοιων αποτελεσμάτων χρησιμοποιώντας προγραμματιστικές διεπαφές από ιστότοπους πολυμέσων. Τεχνολογίες σαν κι αυτήν προϋπήρχαν ως μέρος εφαρμογών όπως iTunes αλλά μόνο πρόσφατα οι εταιρίες που αναπτύσσουν λειτουργικά συστήματα υλοποίησαν μεθόδους αναζήτησης που συνδυάζουν δεδομένα που υπάρχουν στον υπολογιστή και στο διαδίκτυο.

Το κεφάλαιο 5 περιέχει μία μελέτη περίπτωσης αποθήκης δεδομένων. Μελετήθηκαν δύο προβλήματα συγχρονισμού μέσω διαδικτύου: χρονική ολοκλήρωση δεδομένων και εξαρτώμενη ολοκλήρωση δεδομένων. Σε αυτό το κεφάλαιο παρουσιάζονται αναλυτικά τα προβλήματα καθώς και οι λύσεις που υλοποιήθηκαν για την αντιμετώπισή τους. Οι μέθοδοι που αναφέρονται για την επίλυση των προβλημάτων υλοποιήθηκαν και χρησιμοποιούνται μέχρι και σήμερα στο σύστημα που μελετήθηκε.

Στο κεφάλαιο 6 μελετήθηκε η κατάσταση της διαδικτυακής ολοκλήρωσης των δεδομένων στον βιολογικό τομέα που υποφέρει από ασυμβατότητα δεδομένων κυρίως λόγω της ύπαρξης πολλαπλών standards για κάθε βιολογικό τομέα και έλλειψης συμβατότητας end-to-end (από την πηγή των δεδομένων μέχρι τον τελικό χρήστη). Η μελέτη παρουσιάζει την κατάσταση των δεδομένων στον τομέα αυτόν και προτείνει λύσεις ως προς τον ιδανικό σχεδιασμό για την ολοκλήρωση των δεδομένων αυτών.

Στο κεφάλαιο 7 παρουσιάζεται μία εφαρμογή με σκοπό την στοχευμένη αναζήτηση διαδικτυακής διδακτικής ύλης στον τομέα της βιοπληροφορικής. Χρησιμοποιήθηκε η προσωποποιημένη προγραμματιστική διεπαφή της Google για την αναζήτηση και προγραμματίστηκε με σκοπό την υιοθέτηση του συστήματος σε δημοφιλείς ιστότοπους. Η

εφαρμογή αυτή χρησιμοποιείται ήδη από επιστήμονες του τομέα της βιοπληροφορικής και ερευνώνται ενεργά τρόποι υιοθέτησής της σε εταιρίες του τομέα αυτού.

Γενικά είναι φανερό ότι στη σύγχρονη κοινωνία, όσον αφορά την κατανόηση και περαιτέρω ανάπτυξη του διαδικτύου είναι απαραίτητη η διεπιστημονική συνεργασία. Μέσω τέτοιου είδους συνεργασίες είναι δυνατόν να υλοποιηθούν χρήσιμες τεχνολογίες και εφαρμογές που θα βοηθήσουν τους επιστήμονες και ερευνητές να επικεντρωθούν στον σκοπό τους αντί να προσπαθούν να μάθουν πως να χρησιμοποιούν πολύπλοκα εργαλεία. Η παρούσα διατριβή παρουσιάζει μία πληθώρα εφαρμογών, εργαλείων και τεχνικών που μπορούν να χρησιμοποιηθούν για την εξέλιξη του διαδικτύου. Λαμβάνοντας υπ΄ όψιν ότι ένα σημαντικό ποσοστό από τις τεχνολογίες που παρουσιάστηκαν έχουν ήδη υιοθετηθεί από κυρίαρχες εταιρίες του εκάστοτε τομέα, συμπεραίνεται ότι η διατριβή έχει καλύψει επιτυχώς τους στόχους της.

Acknowledgements

I would like to acknowledge Vicky for her help, guidance and support during authoring of this thesis. Also I would like to acknowledge my friends Ilias, Patrick, Vaggelis, Stratos, John and Michalis for their support and friendship throughout all these years.

Abstract

The Web has undoubtedly transformed the way humans live their lives. This is mainly due to the amount of information we can receive with just a few keystrokes. This thesis presents a study on cutting edge web technologies and specifically focuses on data integration. In more detail this thesis examines and demonstrates: Ways to mine and use desktop data to improve user experience on the web; A concept that utilises a number of web technologies for collaborative e-Learning; Enhancements on the desktop search that combine desktop with web results; A case study on a live data warehouse system and how to resolve synchronisation issues on these type of systems; A review on the current state of biological data; and a web application that performs a targeted search to sites related to bioinformatics and retrieves training materials from these websites. These are just a few of the plethora of fields that the web and data integration technologies can prove to be helpful for the advancement of mankind. With the right tools in place and techniques intelligent enough to provide the user with the data needed, the risk of being lost in the huge amount of information that is freely available decreases. Also the processes of learning, technological advancement, information retrieval by using the web infrastructure will be and in many aspects already are part of people's daily lives.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The proliferation of new technologies and internet tools is fundamentally changing the way we live and work. The transformation in which the web has permeated our lives is amazing. Not only in the technology, developments and applications, but also at the sociological and educational levels. It was only 1993, less than 30 years ago that the WWW became a free and available resource to all. A crucial step as underlined here by its creator Tim Bernres-Lee "CERN's decision to make the Web foundations and protocols available on a royalty free basis, and without additional impediments, was crucial to the Web's existence. Without this commitment, the enormous individual and corporate investment in Web technology simply would never have happened, and we wouldn't have the Web today." Another fundamental step towards the ability we have our days to contribute and develop new technologies and applications derives from the efforts of the World Wide Web Consortium (W3C) in 1994, an international community devoted to developing open web standards. The past six years have seen an explosive proliferation in terms of developments, technologies and applications on the web, with advances across multiple platforms and the web increasingly becoming

more mobile and so its users. These developments are also applicable across all possible disciplines and realms of our daily lives.

1.2 Overview

This thesis is about data integration technologies on the World Wide Web. In the beginning it examines and experiments in small scale with a simple application that uses data integration to take advantage of the user's bookmarks in order to be used in the web. Later it applies the knowledge learned from this projects to create an integration concept for e-Learning. Afterwards there is more experimentation with enhancing existing applications and specifically the desktop search by integrating web features. The next step was to get to system level and explore complex integration issues that regard timing and dependencies when integrating data between two large scale systems. Later on, it goes even further to examine the status of the data in a whole field where integration is a paramount issue, namely the field of biological research. Finally, to assist the bioinformaticians with the discoverability of training materials in order to make it easier to advance this field, an application was developed that makes it easy to search and get training materials from popular bioinformatics websites.

1.3 Research Questions

This thesis explores a variety of facets and their relevant applications and solutions related to the web. In particular, the topics of web mining, data integration and web services, and their application in a number of different settings: e-Learning, desktop search, bioinformatics, training and data warehouses.

The following questions and problems are addressed through this thesis:

- Is it possible to use desktop data in conjunction with web technologies in order to make the web more personalised? Currently the trend is to convince the user to store

his data online and this way each company can use them to enable personalisation features. The approach on this thesis respects the user's privacy and just stores minimal information regarding the results of the integration and not the actual data used.

- Using the web to enhance learning experience: how can e-Learning be improved by using advanced web technologies? The concept suggests a number of features that can improve the user experience on collaborative e-Learning. Currently some of the features suggested are part of popular e-Learning platforms and the author is looking forward to start seeing some of the advanced features like automatic background search and "smart" wiki annotation in real world implementations.

- Desktop search: how can the search process on a computer be enhanced with web features to provide the user with relevant data to the ones he is searching for? In the scope of this work an application that is achieving this was created. Recently similar features were added in popular desktop searches.

- Data warehouse issues regarding timing and dependencies: how to resolve these issues based on a real case study. Although the suggested solution contains social features that are still to be adopted by the major technology providers.

- Data integration is still a challenge and there are clearly specific gaps associated to different disciplines and fields. How is the situation in biological research where recent high throughput advances and technologies have brought researchers to a stilt in data storage, analysis and integration? In the scope of this thesis there was an overview on this topic and a solution for the ideal state of data integration is proposed.

- Looking for bioinformatics training materials is a daunting exercise when the typical user "googles" and then visits each link one by one in order to verify the presence of interesting materials. How can this process be improved? Currently there are some efforts that independently try to be centralised sources for bioinfotmatics training

materials. The work of this thesis attempts to integrate these websites by doing a targeted search using an API.

1.4 Research Challenges

Like any other field, in informatics there are various challenges that should be tackled in order to produce a comprehensive research on a given topic. This thesis is no exception. There were many challenging difficulties that had to be resolved that regard both the knowledge base for data integration and also since it is a very technical field, there were technical challenges while developing applications either as proof of concept or improvements to existing systems. The following list describes some of the many challenges that were faced during this thesis.

- **Keeping up to date with latest technologies:** In the field of informatics, programming languages, tools, technologies and techniques evolve very rapidly. As a result it is always a challenge to be up to date with the latest trends and technologies while they are being developed/released.

- **Best practice:** On programming problems, there are usually many solutions to a given issue. Trying to find out which one can be easily adoptable, extendible and re-usable is a major issue in the field. Writing code that can be considered "best practice" is a challenge for every software developer.

- **Innovation:** Innovation is very important in computer science. Coming up with fresh ideas and solutions that have never been applied or even extending or getting a new approach on old ideas are paramount issues that make a huge difference both in the academic and corporate sectors.

- **Terminology:** Similarly to other fields, when it comes to terminology, there are many ways that people defined the same thing. Especially when working on upcoming

technologies, some times it is hard to figure out if two terms are referring to the same subject. Also some terms are getting deprecated and discarded or replaced by other terms.

- **Related work:** Since there is a significant world wide effort on improving/inventing web technologies, sometimes it is very challenging to find related work on the subject that an individual is working on. Some projects/efforts can seem similar but eventually do something totally different or try to do the same thing by taking a completely different approach that required knowledge of other fields/tools/concepts.

1.5 Contribution

This thesis tackles problems and suggests solutions across various disciplines, both academic and corporate. Each chapter provides a comprehensive view of a given problem along with their suggested/implemented solutions. Here is a list of the contribution in the given scientific field per chapter:

- Chapter 2 describes a successful attempt to combine desktop and web data in order to profile a user. Technologies like this are already in people's lives with data mining done by browsers themselves in order to improve a company's services. Similar attempts to unify desktop data on the web was done in projects like MyTag [66] and WikSAR [12] but these projects mainly attempt to just replicate the desktop data on the web.

- Chapter 3 suggests various web technologies to be used for collaborative e-Learning. Today we can see some of them already in the market from popular implementations like blackboard. There are still some advanced features that are not being used yet. Therefore there is still potential for people to get inspired and help make collaborative e-Learning 'smarter' and more flexible.

- Chapter 4 contains a successful attempt to unify desktop and web data in a single search. This projects extends upon the popular search mechanisms that are built in all major operating systems. Lately similar implementations have been observed in major operating systems that enhance the desktop search experience. However these attempts do not include data for multimedia files which is the main feature of the proof of concept application.

- Chapter 5 describes and provides simple solutions to problems regarding timing and dependencies upon syncronisation in data warehouses. When someone encounters similar issues, they can be inspired to create their own implementations according to the solutions suggested. There are other approaches that attempt to solve similar issues [77] but for more complicated systems that required significantly different infrastructure than the system examined in this case study.

- Chapter 6 is a review on data integration in biological research, a scientific field that suffers from data heterogeneity. The chapter investigates the data heterogeneity problem and provides the audience with a view of the ideal state of data integration and suggests how to get there.

- Chapter 7 provides a web application that attempts to focus a Google search to sites that contain training materials for bioinformaticians. The web application has the potential to help educating bioinformaticians around the world. Current attempts to achieve such a thing focus on either relying on the users to provide the materials [10] or they attempt to create tools to parse websites in order to discover what materials are available (TeSS).

1.6 Structure

As one can see, the present thesis covers a wide spectrum of aspects and actual applications in different sectors, from academic to corporate settings. Each chapter of this thesis is devoted to a specific problem related to data integration on the web that become more and more challenging as it progresses. Here is an overview of how each following chapter resolves the questions mentioned above:

Chapter 2 on "Combining desktop data and web services to profile a user" describes a proof of concept application for profiling a user based on the bookmarks saved in their browser. The application can serve as an API (Application Programming Interface) and the results of processing can be easily shared with other applications to enhance the user's experience, for example, to customise web data.

Chapter 3 presents a concept for collaborative e-Learning in the web and how this is based on the ability to combine and integrate Web content and services to improve the end-user experience. This chapter describes an educational concept in order to use web based techniques to achieve daily learning. The suggested web application uses tutoring and collaborative techniques in a Web environment which has the features of synchronous, asynchronous and social learning.

Chapter 4 discusses web enhanced desktop search utilizing data from multimedia social sites. This chapter describes an attempt to embed web services in desktop applications (more specific, the desktop search) and then study the results. This is done by using social media API's to discover related content to the search results.

Chapter 5 is technically more challenging and describes a real problem solving case study dealing with data integration. It proposes a real life implemented solution for resolving synchronization issues on data warehouses by scheduling requests.

Chapter 6 provides a comprehensive review on data integration in biological research. This chapter illustrates the background on what is data integration from a computational

science point of view, how it has been applied to biological research, which key aspects contributed to its success and future directions.

Chapter 7 presents a Bioinformatics Autodiscovery of Training Materials (aka BATMat), an open-source, Google-based, targeted, automatic search tool for training materials related to bioinformatics. BATMat helps gain access with one click to filtered and portable information containing links to existing materials (when present). BATMat also offers functionality to sort results according to source site or title.

1.7 Publications

The work for this thesis lead to a number of publications on international conferences and scientific journals. Following there is a list of the publications related to this thesis:

[c1] Lapatas, V., & Stefanidakis, M. (2010). Combining Desktop Data and Web 3.0 Technologies to Profile a user. In WEBIST (1) (pp. 350-353).

[c2] Giannakos, M., & Lapatas, V. (2010). Towards web 30 concept for collaborative e-learning. In Proceedings of the Multi-Conference on Innovative Developments in ICT (pp. 147-151).

[c3] Lapatas, V., & Stefanidakis, M. (2011). Web Enhanced Desktop Search Utilizing Data from Multimedia Social Sites. In Proceedings of the Fourth International Conference on Internet Technologies and Applications (ITA 11).

[j1] Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., & Schneider, M. V. (2015). Data integration in biological research: an overview. Journal of Biological Research-Thessaloniki, 22(1), 1-16.

[j2] Lapatas, V., & Stefanidakis, M. (2015). BATMat: Bioinformatics Autodiscovery of Training Materials. Briefings in bioinformatics, bbv071.

Chapter 2

Combining Desktop Data and Web Services to Profile a User

2.1 Introduction

Web mining is a key component of the evolution of the web. With web mining, data on the web can be processed by computers for further use. Data used as input source to web mining can be contents of web pages (Web Content Mining), user activity (Web Usage Mining) or website structure data (Web Structure Mining).

All this information resides at server side and is constituted of the data that the user, voluntarily or not, submits to web services' sites. On the other hand, an interesting aspect of emerging Web technologies, currently underdeveloped but with future potential, is the seamless integration of desktop and web data. This chapter aims to show that desktop data residing on user's client side, can be used in the Web to create even more "intelligent" websites.

Projects connected to the social web like MyTag [66] and Semantic Blogging [144] try to unlock the previously underutilized user's data. Other implementations suggest the usage of web technologies in a desktop environment [12, 171]. Those technologies can facilitate

the web and desktop data integration. However, there appears to exist no seamless use of desktop data in the web until recently. Applications that use desktop data on the web always need user interaction and they usually have to store user files on the web.

The following sections show that with present technologies, desktop and web data integration can be achieved. As a proof of concept, an application was created that can profile a user based on the bookmarks saved in his browser. The application can serve as an API (Application Programming Interface) and the results of processing can be easily shared with other applications to enhance the user's experience, for example, to customise web data.

The rest of the chapter is structured as follows: in the second section related projects are being presented and discussed. The third section describes the application and presents some results concerning its accuracy. Privacy issues are addressed next and the final section lists future development ideas along with the author's conclusions.

2.2 Related Work

Over the years a few similar ideas via different approaches were presented, summarised in this section.

MyTag [66] presents user's data that are stored on the web as a cross web-based interface. MyTag exploits web services from various sites to access user data. The application presented in this chapter is based on a similar idea but uses desktop data for user profiling and needs no user interaction.

Semantic Blogging [144] uses desktop data for the needs of blogging activities. With Semantic Blogging a user can easily handle his own desktop data (contacts, calendars etc.). Semantic Blogging is the most closely related application to the proposed approach but again needs user interaction in order to be used.

WikSAR [12] is a web application that uses desktop data in a wiki environment (such as addresses, calendars etc). Although this project can use local data on the web, it doesn't process them by any means; it only presents them with a more elegant way of browsing.

Gnowsis Semantic Desktop [171] translates desktop data into semantic data for major operating systems. Although not web related, it supplies an easy way to access desktop data based on semantic analysis.

Automatic Bookmark Classification [23]: This project has a lot in common with the one presented here, as it also classifies a user's bookmarks. However, there is no automated usage of the classification process; the user is prompted to accept the classification result or insert the results he believes that suit best.

Personalized Search [188] studies the impact of web mining techniques on a search engine. They use of logs from previous searches as well as previously visited web sites to profile a user and return more accurate results. Only web data is being used, not taking advantage of desktop local data.

2.3 The Application

In the research line of this thesis, the aforementioned idea of exploiting desktop data on the web, was explored through an application has been developed using the Python programming language. This application is able to profile a user by using the bookmarks he has assigned to his web browser. This program has no user interface and the results are being acquired without any user interaction.

This application takes advantage of web services in order to detect what the user is really interested in. The web service being used is a classification service [193]. URLclassifier is a web service accepting a URL and returning the categories that are embedded in the relevant web page, as a result. After the application gathers the user's bookmarks, it uses this service

to classify them into categories. Those categories are closely related to the user's fields of interest and can be used to profile the individual whose bookmarks were processed.

An on-line classification service is being used in order the local application not to consume a lot of processing power. Thus, the application can be used in mobile devices that do not have the processing power needed for the execution of a classification program. With the on-line service, full text of web pages is being processed. Another way of creating a light-weight classifier is to use only the URL of a web page as input to classification [19, 112].

The application extracts the user's bookmarks and processes them with the help of the on-line classifier. Each of the popular browsers has a different way of storing its bookmarks. Safari stores its bookmarks in a .plist file while Firefox uses a .html file and Opera a plain text file. For everyone of the above browsers a different parser was developed, enabling the program's first phase to gather bookmarks from each one them. In a subsequent phase, the application sends the gathered bookmarks one by one as input to URLClassifier, in order to get the necessary results.

After the classification, all the results are stored into a list which can be consumed by a third party application or just be printed for debugging purposes.

2.3.1 Application's Accuracy

A number of computer science students was asked to submit their bookmarks to use them as input for the presented application, in order to test the application's accuracy with real life samples. Some samples contained a big number of bookmarks (greater than 100), while some others only few (less than 10). This was advantageous to the experimentation because it was possible to examine if the success percentage of the application varies between different input sizes.

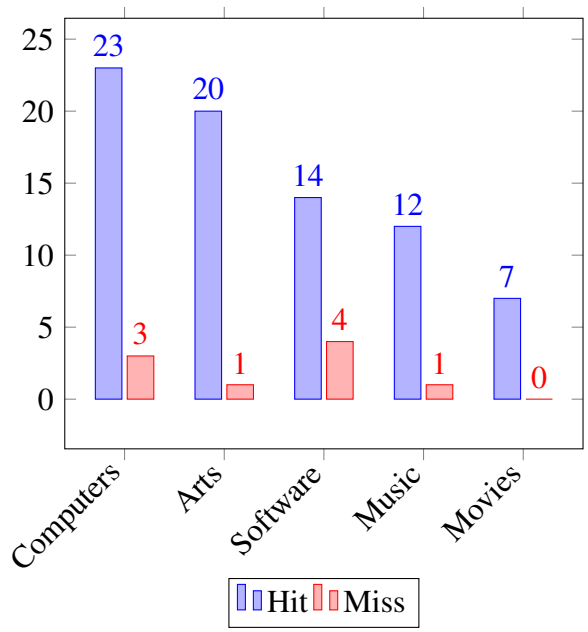

Figure 2.1 Top five results of the biggest set

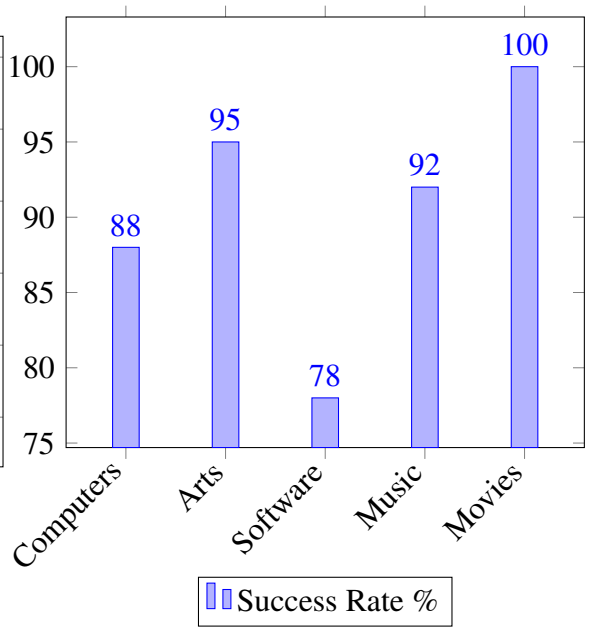

Figure 2.2 Success rate of the biggest set's top five results

After applying the classification process that was described in the previous section, the outcome was handed back to the users in order to evaluate the correctness of the classification results. From their answers a set of hit-miss counts was generated.

Figure 2.1 shows the five results of a single set with the maximum number of bookmarks. This particular set consists of 120 web sites and it was the largest set of those that were tested. The left bars indicate the successful categorisation of the sites while the right bars indicate the false categorisation of the sites. As someone can see in this set, the classification is quite accurate. To support this speculation, Figure 2.2 shows the success rate of the top five categories with an average success rate of 89%. The total success rate of the whole sample (including the omitted results) is 72%.

The application presented here works only for sites in English because URLclassifier service does not have support for other languages. Sites that are not in English are being ignored for now. In this particular set around 50 of the 120 sites didn't return any results and almost all of them were sites with non-English text (41% of the sample).

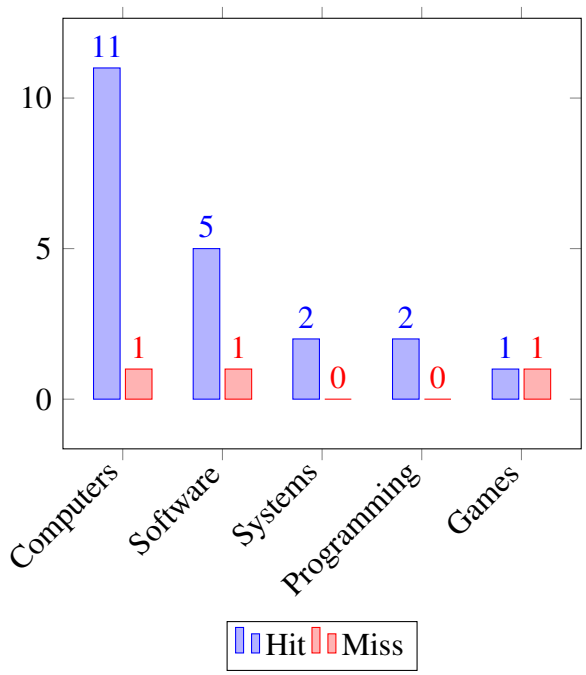

Figure 2.3 Top five results of the smallest Set

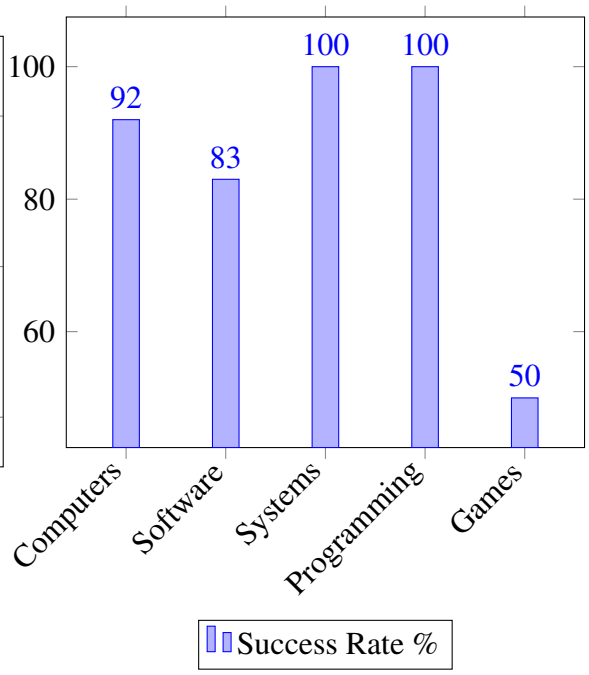

Figure 2.4 Success rate of the smallest set's top five results

Let's take a look now to a complete different set, which consists of 22 bookmarks and was the smallest set available. Once again, the results were quite accurate (Figures 2.3 and 2.4). The percentage of the success in the top five categories is 87,5% while the total success rate of the set is 87%. In this case those two values have a very small difference because almost all of the results are included in the top five categories (only 7 results are being omitted).

Like the first set, this one had also some sites that didn't return results because those sites were using the non-English language. (7 sites or 31% of the sample).

The figures shown before indicate that the classification success rate and the relevant user satisfaction are high. When the users were asked about the results of this application, most of the feedback was strongly positive and even the worst review was also positive.

2.4 Privacy Issues

In order to respect the user's privacy, whenever classification data is going to be sent to a Web service, it is absolutely necessary not to allow the presented application to leak information about any of the user's bookmarks or personal files. The only information that should pass from the user to the server will be the classification results.

Even if classification only results are allowed to leave the user's computer, this action can be allowed only after user's agreement. A possible solution can be a license agreement. Unfortunately, the problem with license agreements is that very few people actually read them.

In order to increase user's awareness of the application actions and make sure that he understands completely what this application will do in his machine, another method is planned to be used. The first time the application is executed, an intuitive configuration window will be used to give the user the chance to review and control the data that will be processed, instead of just displaying incomprehensible license terminology. The configuration dialogue will appear only once and thereafter the application will be able to run transparently, without any further user interaction. To ensure the existence of a constant but at the same time unobtrusive warning to the user, small notification icons or messages, that do not require user response, may be used at the desktop.

Private data leakage prevention is a major issue in all desktop-web integration efforts and not specific to the application presented herein [58, 190]. It is obvious that a total solution provided by standard OS services is needed, as desktop and web spaces are getting fused within each other.

2.5 Conclusion

This chapter tackles a simple data integration problem by investigating ways to utilise private data like user's bookmarks, locked until recently inside a user's computer for the intelligent generation of Web content. An application, which generates the profile of a certain user based on the content of his browser bookmarks, was demonstrated. An external classifier service enables the execution of the application on platforms with limited processing capabilities like mobile devices and netbooks. The resulting figures indicate that with the usage of a generic non-trainable and non fine-tunable classifier it is possible to achieve satisfactory results in over 70% of cases in average and near 90% for top 5 categories in user's profile. The knowledge that was gained from the study and development of this application, was expanded further to create a concept for collaborative e-Learning that takes advantage of the latest trends in web technologies.

Chapter 3

Towards a Concept of Utilising Advanced Web Technologies for Collaborative e-Learning

3.1 Introduction

Nowadays, one of the hottest topics in education is the opportunities that advanced Web technologies can offer by handling the WWW as the largest information database humans have ever invented. People can access large amounts of information (e.g. news, research etc.) with just a few clicks of the mouse by using automated personally configured search engines without even knowing it.

To get to this point the WWW had to be evolved from text-based static pages. More specifically the "first version" of the Web (Web 1.0) introduced great opportunities in open and distance learning. It was the first time in human's history where the tutor could transfer educational content to the learner by using easy-to-access, visualised techniques.

Later, with the transformation of Web 1.0 to Web 2.0 the WWW gained a vast of new features and soon enough web-sites/applications like wikis, blogs and social networks became a part of most people's lives.

In education, the actual contribution of Web 2.0 lies in the learner's ability to be able to interact with web content. As a result it enables the learner to add comments, reply or even change information created by his tutor, instead of passively reading it.

One step further from Web 2.0, the "advanced" Web can be considered as the sum of the efforts that are currently being made to make the web "smarter". One of its most important features is the ability to combine and integrate Web content and services to improve the end-user experience.

Inspired by the technologies that are emerging in the Web, an educational concept was conceived and designed in order to use those techniques to achieve daily learning. The suggested web application will be using tutoring and collaborative techniques in a Web environment which will have the features of synchronous, asynchronous and social learning.

The rest of this chapter is structured as follows. In the second section, Web 1.0 and Web 2.0 technologies are being analysed, as well as their contribution in education. In the third section the abilities of the advanced Web and what they can contribute in the learning field is being presented. In section number four, this chapter's concept is being described. Furthermore, possible ways of implementation, both in corporative and educational environments, are being discussed. In section five, as future work, technical details of ways that this concept can be developed are being mentioned. To conclude, in the final section, the possible impact that an application like the presented one can have in education is being discussed.

3.2 About Web 1.0 and Web 2.0 in Education

Web 1.0 is the term used to refer to web in the form existing from 1990 to 2000 [155]. It allowed data sharing over the Internet. The Web 1.0 was divided into working directories;

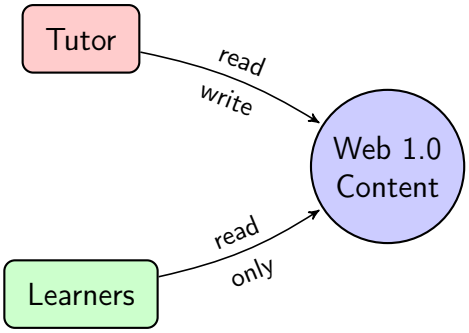

Figure 3.1 Web 1.0 capabilities in education

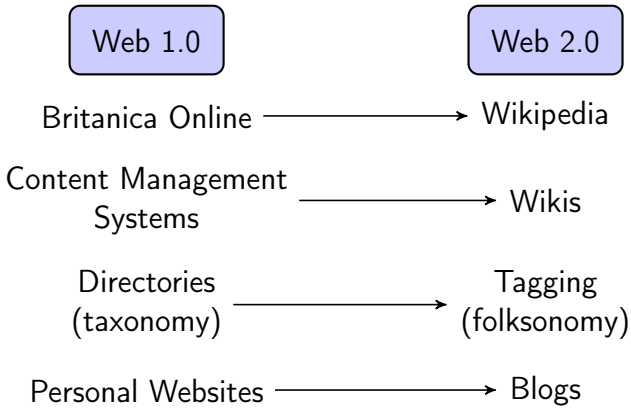

Figure 3.2 The evolution of the websites

practically everyone had their own space [48]. For educational purposes Web 1.0 provides the technology platforms which publish knowledge content. But the limitations on content creation limited the potential. The techno-centric nature of Web 1.0 could not satisfy educational needs. For that reason Web 1.0 educational usage was limited to publish content (Figure 3.1).

"Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform. Chief among those rules is this: Build applications that harness network effects to get better the more people use them" [63]. The read/write Web 2.0 consists of a set of new technologies

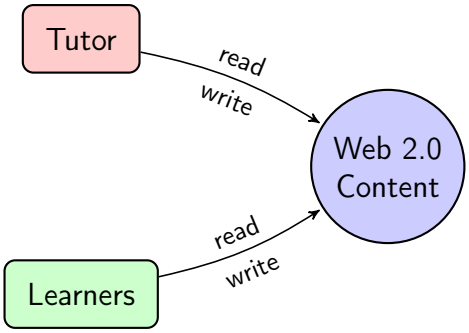

Figure 3.3 Web 2.0 capabilities in education

that make the web more like a platform. With the evolution of Web 1.0 to Web 2.0, there was a transformation of the applications (Figure 3.2).

Transformation also affects the area of education. Applications like E-learning 2.0, Classroom 2.0 and Enterprise 2.0 appear [141]. These applications pay attention in the user's ability to interact and manipulate the educational content (Figure 3.3).

3.3 About Advanced Web Technologies in Education

The advanced Web is the third stage of the web evolution (Figure 3.4), that began recently [25]. The current trend is to transform the World wide web in a massive, open and queryable database along with Application Programming Interfaces (APIs) that can intelligently handle these data. This happens towards making the Web more "smart" and personalised.

In the past decade, education concepts based on Intelligent Tutoring Systems (ITS) [36] and advanced Web technologies have been mentioned [156, 157]. The rapid evolution of learning software, artificial intelligence and web technologies make ITS and the advanced Web a viable option. Moreover, the advanced Web offers more intelligent services and in addition to reading and writing content, user's actions can initiate web processes (Figure 3.5), that can be possible with technologies like smart interfaces and intelligent agents.

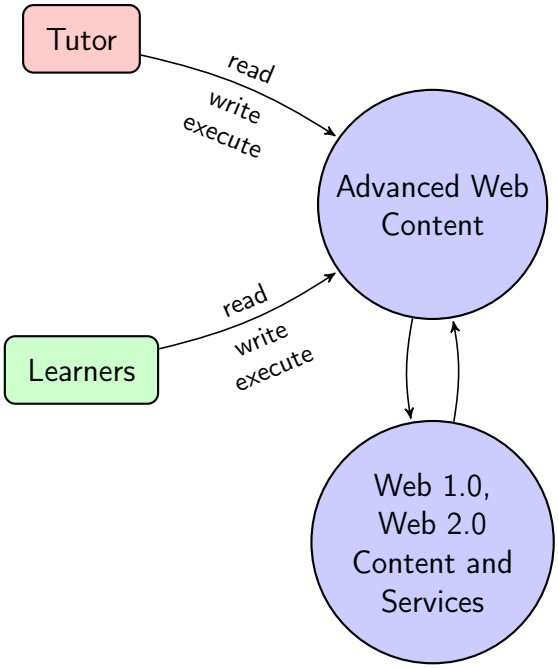

Figure 3.4 Web 3.0 capabilities in education

3.4 Our Concept

Our concept consists of a suggestion interface to improve the e-Learning experience using advanced Web technologies. An interface like that can enhance collaborative learning with smart interfaces and auto-updated content depending on the topics of discussion. Its most important features can be:

- Synchronous/asynchronous and social content
- Wiki-enabled interface
- Customised background search

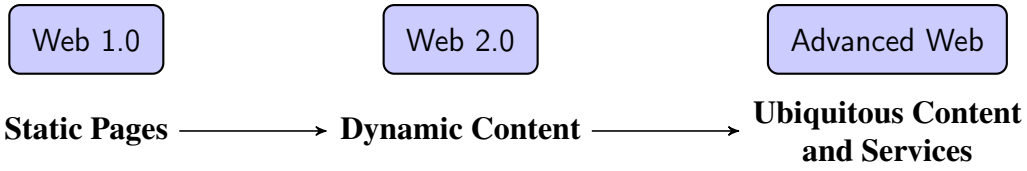

Figure 3.5 The evolution of content and technologies on the web

3.4.1 Synchronous/Asynchronous and Social Content

The website will contain on-line chat (both text and video) to enable communication between tutors and learners, as well as between learners. This way everyone can cooperate or answer questions between them.

This content can be live (synchronous), meaning that when a person interacts with the interface to add content, everyone is able to see the changes in real- time and there is no need to refresh the page in the browser. Also, the content can be asynchronous, meaning that if someone wasn't able to be on-line when a change took place, he will be able to review the changes any time he logs on the site.

To make this more understandable a usage example is presented (Figure 3.6). The tutor creates a new subject in the site and then he logs off. Then, some learners enter the site and see the new subject. One of the learners has a question to ask and submits it. The learners initiate a discussion, live, in attempt to solve the problem.

Then after a few hours the tutor logs in and sees the learners' conversation. He notices that the subject is too abstract and decides to change it so that it will be less confusing.

Then he notices that some learners are on-line at the time and they start a video conference to solve their questions. After the video conference comes to an end, another learner logs in and sees the changed subject title and that he missed the video conference. This is not a problem because he can stream the video conference to his computer and see what he had missed.

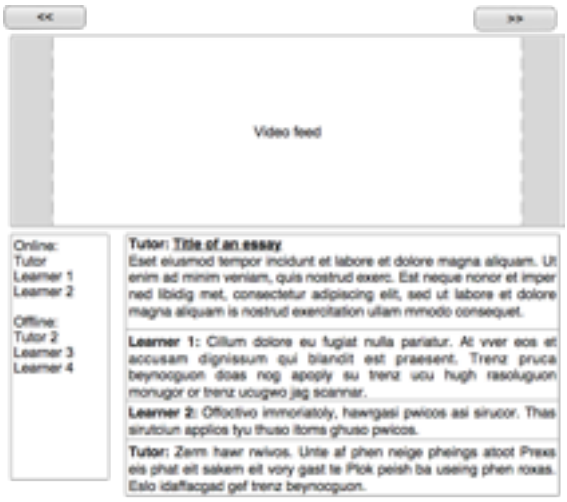

Figure 3.6 Example of a low fidelity prototype using synchronous/asynchronous and social content

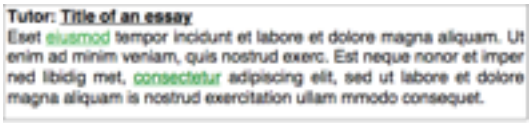

Figure 3.7 A hypothetical wiki-enabled interface

3.4.2 Wiki-enabled Interface

Since the most common reference library in the world is wikipedia.org, it would be more than necessary to embed an interface which links an e-Learning site with wikis (Figure 3.7). To do so, there are two possible ways.

The first way, and the most obvious one, is to let the tutors define which words of the text will be linked to wikis. This is practical but it makes the work of the tutors much more complicated.

The second way is to do this automatically. The text in the conversations can be processed and let the on-line application decide which words can be wiki- enabled.

In this part of the project instead of traditional wikis it might be useful to add visual wikis [92] to enrich the page's content. This feature will offer a more attractive interface for the end user.

Figure 3.8 A low fidelity prototype of how the whole web- site could be. The search results are being displayed on the right

3.4.3 Customised Background Search

The most important feature of this web application is the customized background search.

This could be a Google search Application Programming Interface (API) which performs searches to some default educational sites and the tutor can also add more websites if he likes to.

The interface of this search engine will be invisible and the arguments that are going to be used will be the content of the conversations as well as the subject titles.

The top hits of the search results will be displayed in a reserved area in the web interface. Those results will vary depending on the content of the conversation. So for example if some people have a live discussion about a subject, automatically some web-sites will be suggested to them in real time as they communicate with each other.

Another way of implementing such a service is to use ontology mining techniques [104, 186] in specific web sites. In the actual implementation of this concept both techniques will be tested and finally, the one returning the most appropriate results will be used.

This could be really useful for people who can not use search engines effectively and some results could be also helpful for easy access to related websites.

3.5 Future Work

The concept of e-Learning via advanced web technologies led to the design of the proposed architecture. Since a system like this is not close to get introduced to the public yet, an implementation attempt should go through a development and testing circle based on the following.

At the beginning, there will be a need of a low fidelity implementation of the proposed system. This will be an assessment of the resulting enhancement in the educational process and lifelong characteristics [169] induced by the usage of the proposed system.

One important factor, following a possibly low fidelity implementation of the proposed system, will be an assessment of the resulting enhancement in the educational process and lifelong characteristics [169] induced by the usage of the proposed system. At first a desktop application will be developed for testing purposes. This technique will be used because the development is much easier by using the system's API's and frameworks to execute otherwise complicated actions. This concept will be evaluated in contrast with other effective learning environments in between-group experiments. If the evaluation results are negative, the application will be redesigned and redeveloped until positive feedback is received. In that case, a highest fidelity application will be developed in order to be ready in its web form.

Even thought, today's web technologies enable personalization [110], an attempt to increase it through interactivity will take place. This can be accomplished by using cameras and microphones in conjunction with gesture, facial and sound recognition algorithms [195]. Those will extend the system capabilities by enabling emotional recognition in order to offer more personalized feedback. In addition, profiling techniques based on the user's data [126] can be used to achieve more accurate personalization.

The web development can be achieved by using AJAX (Asynchronous JAvascript and XML) technologies which provide all the necessary tools needed for the proposed features, namely: synchronous/asynchronous conversations, video conferences, links to wikis and the

customised search. AJAX technologies can enable the manipulation of the website content in order to use it as input for a wiki site or a customised Google search, return the results and finally process them to create some output.

This is just a first assumption of how this project can be implemented using today's technologies. Of course the above mentioned technologies are subject to change according to what is considered "cutting edge" at the time of the implementation.

3.6 Conclusion

Technologies such as Artificial Intelligence and the WWW have rapidly evolved over the last few years. Despite of this situation, no educational applications that utilize fully the potential of advanced web technologies have been developed. The goal of this concept is to introduce the user to lifelong learning. The daily usage of the application comes from its social characteristics and the adapted content which motivates the user. Using the combination of the features mentioned above (synchronous/asynchronous conversations, video conferences, data integration for wiki-enabled interface and automated background search), the system will be able to achieve a personalized interactivity with each user. The presented concept is an attempt to introduce an educational system which combines advanced Web technologies in order to achieve better personalization and usability. Furthermore, the social networking characteristics will contribute in gaining wide acceptance and satisfaction.

Summarizing, the proposed system could be a useful medium for rapid and accurate knowledge spread in academic and corporate sectors. During the research phase for a suitable e-learning architecture, it was early realized that merging data of various sources increases the benefits for the user of applications. In the same spirit, the next step was to test the integration of diverse data sources to enhance the desktop environment. The outcomes of this effort are presented in next chapter.

Chapter 4

Web Enhanced Desktop Search Utilizing Data from Multimedia Social Sites

4.1 Introduction

Today's trend on the Internet is to port desktop programs on the web [127, 209]. Major technology pioneers like Google Inc., Apple Computers Inc. and Adobe Systems Inc. have created on-line programs (Google Docs, iWork.com, Photoshop.com) that provide the same or similar functionality as the desktop alternatives. This trend, however, goes both ways.

Desktop applications are being enhanced with Internet features to provide additional services and content to the end-user [66, 144]. Unfortunately, most of those applications don't take advantage of a critical field of the web, which is the various on-line services of social networking sites. These services are very important in the evolution of the web and they are one of the main reasons of Web 2.0's success. They are very easy to use and the user can find important information within seconds if he knows where to look. Programmatically too, someone can take advantage of these sites either by using the website's APIs (Application Programming Interfaces) or via normal URL (Uniform Resource Locator) accesses. This is a very practical way of embedding additional functionality to any application

by exploiting the rich datasets of social networking sites, datasets that are generated via powerful recommendation algorithms from vast amounts of user interaction.

This chapter explores ways of embedding web services in desktop applications (more specifically, the desktop search) and then studies the accuracy of the resulting data that were retrieved from the web services. It uses HTTP requests to take advantage of data and services that exist in remote websites. This process takes advantage of the social networking recommendation capabilities of popular web sites (e.g. "people who liked this also like...") without the need to create the recommender functionality from scratch.

The rest of the chapter is structured as follows. The next section is a report to well-known projects merging desktop and web search, that relate to the presented one. Section 4.3 presents the detailed description of this chapter's application. In section 4.4, the evaluation results are being presented. The final section discusses the conclusions from the evaluation results and what there is to explore as future work regarding this project.

4.2 Related Work

4.2.1 Desktop Search

The proposed solution presents an attempt to enhance desktop search with web features. In order to clarify the reason why this work is different from existing projects, there is a need to reference the most relevant ones and explain the features that distinguish them from the presented solution. Since this chapter's application constitutes a desktop search tool, it seems logical to refer to the most well known desktop searches [5, 38, 142].

Over the years there have been a lot of attempts by world leading technology companies and researchers to make the desktop search faster and more efficient. Nowadays, almost every major operating system has a build-in search engine to locate files very fast and accurate. This can happen by creating indexes of the files stored in a computer and then instead of

searching the whole hard drive for the user's request, the search engine searches the index files instead. But until recently, none of those engines had a build-in features to expand the search to the Web. The presented application uses the results of the desktop search to expand the user's search on the Web and on-line services to gather additional results from multimedia social websites.

4.2.2 Search with Embedded Web Features

As for the integration of web features in a desktop search, there is some work done by Google Inc. with their product Google Desktop Search [80]. Regarding the functionality of the presented project, the closest work to that has been done by Apple Inc. and their product iTunes [4].

Google Desktop Search

Google Inc. enhances the desktop search with Web search features [80]. Their approach utilizes a keyword-based Google searcher build in the desktop search tool. This is a practical way to search both the Internet and the desktop for the same keywords but it may not target directly social sites specific to the media type of the desktop results. The solution presented in this chapter has a media content-type oriented way of searching the Internet. For example, if the results returned by desktop search are music files, additional song results from a musical on- line service will be fetched, while, if movie files are found on desktop, then the results will be similar movie titles from a relevant site.

iTunes

The last few years, iTunes have offered an interesting feature. When the user selects a song, suggestions from their online store appear on a sidebar. These suggestions are similar to the selected song but the user doesn't have them in his music library. This is a great feature

and the iTunes store owes a lot of its success on it. This was also the feature that inspired the work presented in this chapter. The proposed idea is to use the desktop search to find not only music but any file the user wants and then acquire from suitable on-line services similar items for music, movies, applications, websites, even documents, according to the file's media type. At this point, the demonstrated application can return music and movie results, while additional media types can be handled by plug-in modules.

4.3 The Application

The presented application attempts to integrate information from social networking and media content recommendation web sites into a desktop search application. This will help the users to find content on the Internet, relevant to what they are searching for on their desktop. This application's graphical user interface (GUI) consists of two tables, a search field, a rating system and, for evaluation purposes only, a button, as shown in Figure 4.1. The user can use the search field to search for files in his computer. Then, he can select a file from the results. If the selected file is music or movie file, the application populates the second table with similar results acquired from media content social networking web sites. This is all the functionality that this program was intended to have. For test purposes, a five-star rating system was embedded and a button to send the ratings to a server in order to enable online collection of user evaluation results for further study.

4.3.1 Conceptual Architecture

In Figure 4.2 is a graphical representation of this chapter's application data flow. The application has a search field for the user's files and uses the system's search engine to return the results inside a table. Then, the user can select any of the files returned. If the selected file is a music or movie file, the program initiates a background search to a web site and then

Figure 4.1 Application window

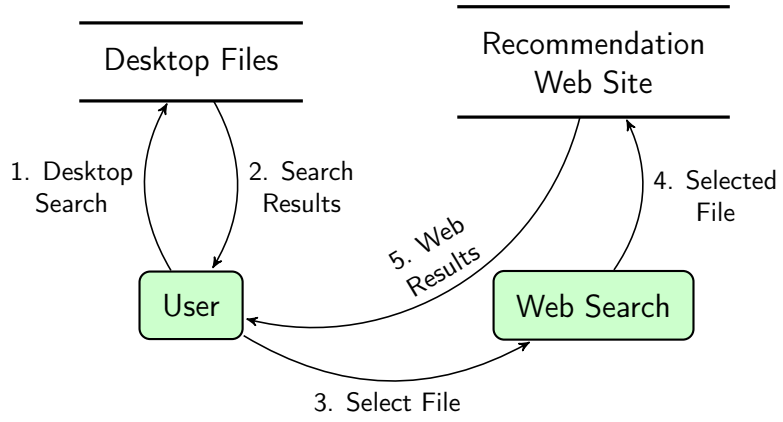

Figure 4.2 Application's data flow diagram

returns relevant results that have been found in that particular site. When those results return to the application, they are being displayed in a second table.

4.3.2 Implementation

The application was implemented under the Mac OS X 10.6 environment, using Xcode 3.2.3 and the Cocoa windows system. The programming language used for the user's application is Objective C 2.0, the data stored to the server in XML (eXtensible Markup Language) format and the collection and analysis of users' evaluation results was accomplished by developing two additional programs in Python.

The Cocoa Framework is used to take advantage of the system's APIs (Application Programming Interfaces) like the Spotlight engine, which is the default desktop search engine of the computers running Mac OS X.

In order to parse web sites and send FTP (File Transfer Protocol) requests to the result gathering server, additional 3rd party frameworks where used. For parsing HTML (Hyper Text Markup Language) web sites an open source Objective C HTML parser was used created by Ben Reeves. The FTP handling was done using the open source S7FTPRequest project created by Aleks Nesterow.

The web sites that this application uses to gather its results are last.fm (http://www.last.fm) for music recommendations and movies.com (http://www.movies.com) for movie recommendations.

4.4 Evaluation

The final application was handed to a group of users to evaluate the relevancy of the suggested results from the web. This was done in order to measure the program's usefulness, as well as, to consider future application features.

For the evaluation procedure, the application was equipped with a typical five-star rating system. This enabled the users to rate each of the additional web suggestion results individually. Each user was responsible to rate the results and send them to a result-collecting server. In the end, the results where gathered together and studied. Some of the users also gave written or verbal feedback of their experience with the application.

Evaluation via rating was used for several reasons. Rating is especially useful in user-based evaluation systems that include ranked items retrieved from the web. Moreover, the rating system goes beyond binary relevancy [206, 210] and fits better to human understanding of multi- grade usefulness.

Another classic approach of user-based application evaluation consists of asking the users to fill questionnaires. This method generally can produce sufficient results for other application types but in this case, since a part of the returned results can be uninteresting to the end-user, this approach cannot be used efficiently. This happens because certain aspects of the program can be rated only as a whole. For example, if the questionnaire asked "how efficient where all the suggested results returned by the application for the song title hysteria", the end user can't answer clearly about the relevance of the recommendation results because there is a great chance that he won't be familiar with a lot of them.

Another well-known approach to evaluate recommendations is to use fixed sets of data where the tester knows the answer set beforehand. This method isn't applicable to this project because the results are being returned dynamically from external web sites and there is no way to compile such evaluation sets.

4.4.1 Music Results

Figure 4.3 shows a chart of the music recommendation results. As someone can deduce from this chart, a total of 3315 songs have been rated. Those where the songs recommended by the application. The users searched for a total number of 70 songs. It is clear that there are very few completely irrelevant results and a lot of extremely relevant results. This means that the users where satisfied from the system and agreed with it's results.

However there where some complains on the feedback that sometimes the application couldn't return any results. This happened somewhere around 1 song every 20 tries. The problem was in either the file name (e.g. it had some illegal characters that weren't removed programmatically) or the web site couldn't find the selected song (e.g. they were non-English language songs).

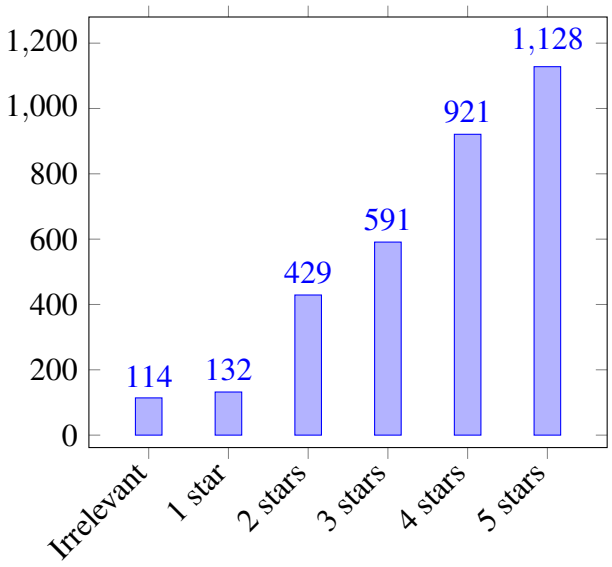

Figure 4.3 Music ratings

4.4.2 Movie Results

Figure 4.4 shows the results of the movie recommendations. The quantity of those results is significantly smaller than the music results because the application was returning in average ten results per selected movie (in contrast to music searches that returned at least 100 results per selected song). Here the selected social web service returned only a small number of results. A different source could increase the number of those results.

According to the data collected, 89 movies have been rated and, according again to the user feedback, approximately one in four movies didn't return any results. Again, this appears to be a characteristic of the web source used. In the movie case there were no file name related problems, as the users where advised to rename their movie files with the proper titles.

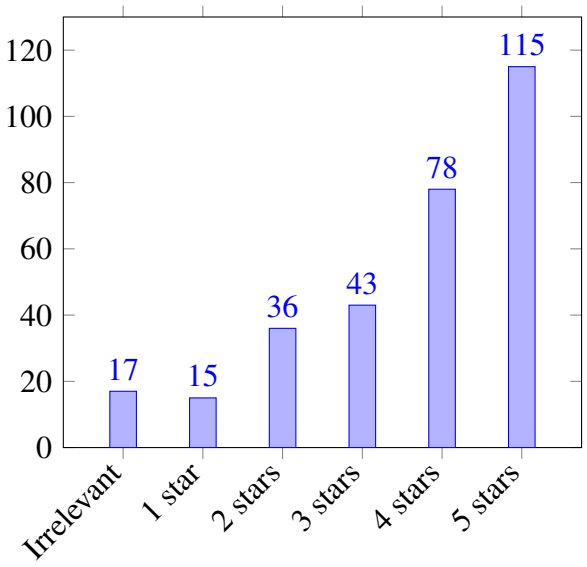

Figure 4.4 Movie ratings

4.5 Conclusion & Future Work

The test result study showed that the users where overall satisfied with the relevancy of the web suggestions returned by the application. This proves that combining desktop data with web sources form media social networking services can bring powerful and trustworthy features to desktop computing. Moreover the application achieves its goal while protecting the user's privacy [58] and without leaking any personal information about the user's identity or files stored within his computer [190]. All website-related actions use simple URL accesses to achieve their purpose. This means that even if the web sites use web-mining technologies, it appears to them as if the user entered their site and searched for a keyword. Therefore, there is no way to determine whether those requests are actually filenames or not.

Some other important results from the users' feedback are that, although most of the users where very satisfied and impressed about the music recommendations, they expected more from the movie recommendation system. Overall the application and the results where better than the author had expected during the building phase of the project, as only two external sources are being used.

There are many possible extra features which can be aimed to increase the relevancy of the results and/or provide additional functionality for the application. Some of the features the authors have in mind are:

- Support for more data types (like web site or document files).
- Cross-reference several sites to find the results to return. This will require more than one resource site for a given file type.
- Post-filtering of the web results. This can be achieved in conjunction with the cross referencing by adding weights to the results based on either the order or the number of websites that returned the same item or both.

The next chapter investigates further the issues of data integration, this time to a complex distributed system that is widely used as the backbone of a plethora commercial websites. More specifically, the issues that were encountered regard the actual synchronization procedure of the data and cases where it was failing due to incorrect timing or dependency factors.

Chapter 5

Data Integration: Resolving Synchronisation Issues on Data Warehouses by Scheduling Requests

5.1 Introduction

Data integration can be described as the process of combining data residing at different sources and providing the user with a unified view of these data [87, 100, 129, 192].

For the past three decades, researchers have investigated methodologies and technologies for data integration systems [86]. Data integration systems always consist of sources that provide the data and a global schema that is used to present a view of the integrated data to the user. Historically, data integration systems have been developed by using two methodologies that are being referred as eager and lazy approach [202].

The eager approach of integrating data is to copy the physical data over to the global schema in order to provide a unified view of these data. This methodology lead to the formation of data warehouses and Business Performance Management (BPM) systems [77].

Systems like these are being used extensively in the business sector in order to help with the decision making process by providing accurate reports of the integrated data.

The main logic behind the lazy approach of integrating data is to keep the physical data on the source schemas and use the global schema as a mapping mechanism in order to provide a unified view of the data. Systems that are using this methodology started with an approach based on Global-as-View (GAV) and then evolved to more complicated systems that are using Local-as-View (LAV) [86, 87, 129, 132, 192] leading to the formulation of GLAV (combination of GAV and LAV) [68, 86] and data exchange systems [86, 123].

Both methods of integrating data have to face challenges when it comes to the integration and the query optimisation process. Regarding the eager method of integrating data, research focuses on methodologies that guarantee data integrity [85, 211] and detailed comparisons of existing applications [41, 173]. Whereas on the lazy method research focuses primarily on two aspects: ways to increase performance and results on the answering query problem [1, 37, 82, 87, 102, 131] and on dealing with data source completeness issues [1, 61, 131, 194].

With the emerge of technologies that enable data integration over the Internet like Linked Data [26, 27, 89], researchers focused their efforts in improving this type of data integration with applications, among others, on e-Science [145], environmental science [207] and bioinformatics [177].

Nowadays, the traditional approaches of data integration like data warehousing have been adopted as practical implementations by several enterprises. However, challenges associated to a variety of aspects ranging from data heterogeneity, query optimisation to timing dependencies remain.

This chapter presents a research conducted on a typical data integration situation and a case study on a data warehouse integration system and more specifically, it highlights two synchronisation issues that were encountered upon development of a real life integration system. Firstly the “timing issue”, a case where the timing in which the data are being

integrated affects their interpretation on the global schema. Secondly the “dependency issue”, which refers to the situation where the source schema is dependent on the global schema for integrating a chunk of data. Meaning that data cannot be synced over to the global schema if a specific action has not taken place.

The rest of the chapter is structured as follows. Section 5.2 presents related research work that has been done on similar issues. Section 5.3 provides the theoretical framework used throughout the solutions presented herein. Section 5.4 provides the system specification along with an abstracted real world example. Section 5.5 describes the timing issue that was investigated along with the implemented solution. Section 5.6 describes the dependency issue for this data integration system along with the implemented solution. Lastly, section 5.7 lists the conclusions of this case study.

5.2 Related Work

Making sure that the synchronisation process is being done in the proper timing on data warehouses has been dealt with before from the scientific community. In more detail, researchers have provided methodologies for dealing with frequent updates/deletes on the source that may cause corrupted data on the global schema [211]. This differs from this case study on the fact that even though we are dealing with timing issues, they are not caused by the frequency of the changes that occur on the source. The problem addressed here is more related to the interpretation of the data on the global schema.

Furthermore, it has been suggested to use a “Right Time Integrator” to integrate data on BPM systems on the proper time [77]. This component could possibly resolve issues like the ones presented here but compared to the herein proposed solution, it is embedded in a much more complex setting.

5.3 Theoretical Framework

In this section we define the theoretical framework based on the work of Lenzerini (2002) [129]. This framework is used throughout the rest of this chapter in order to provide a logical view of the operations that are being presented. Originally this work has been developed to describe lazy data integration systems but it can easily be adopted for use on eager ways too.

5.3.1 Existing Notations

To describe a data integration system I, it is commonly defined it as a triple $\langle G, S, M \rangle$, where

- G is the global schema
- S is the source schema
- M is the mapping between G and S

For a data warehouse system we consider G as the schema that the integrated data reside whereas S is a single source of data.

Regarding the mappings, M can be represented as:

- $q_S \rightsquigarrow q_G$
- $q_G \rightsquigarrow q_S$

This notation describes a successful mapping between S and G. Alternatively, this means a query over S (q_S) and another query over G (q_G) are of the same arity and expected to return the same results. On a data warehouse system we can consider this as the fact that the data are successfully integrated and both q_S and q_G return the same results.

5.3.2 Introduced Notations

In the scope of this work we define:

- $q_S \nrightarrow q_G$
- $q_G \nrightarrow q_S$

as a corrupted data "mapping", meaning that q_S and q_G are queries that should be of the same arity but fail. In other words, q_G will return different results than q_S even though these queries should be equivalent.

Furthermore, we define the message that is being sent from the source to the global schema in order to create the mapping as m_i where i is the index of the message.

5.4 System Specification

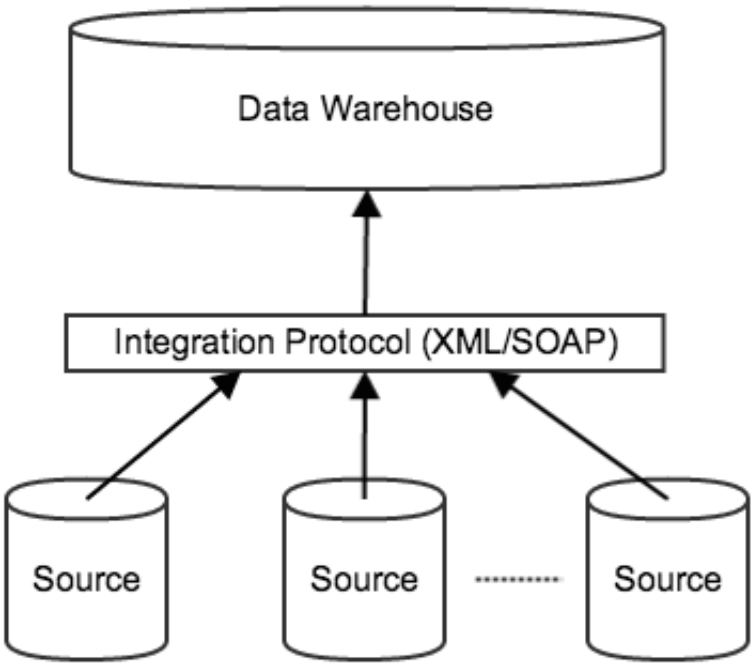

Figure 5.1 The System

The system consists of multiple sources and one global schema using the data warehouse model for integrating the data. In this case study we investigate what happens when an

integration attempt from one source to the global schema occurs. As an example, an abstracted version of the real system is being used since we are not allowed to reveal information about sensitive company data. The source in question is the database of a commercial website and the global schema contains data across many similar commercial websites.

Figure 5.1 shows a simplified version of the data integration system. It worths mentioning that to integrate the data, the sources are required to send their data over to the global schema by using protocoled requests like XML [31] or SOAP [29].

5.5 Timing Issue

5.5.1 The Problem

The timing issue arose after attempting to integrate the source with the global schema. This issue appeared to affect the interpretation of the data in the global schema. The reason for this was due to the fact that the two systems were using different concepts for handling their data. To make this clearer, if we look back to the example, the issue was being caused by“delayed” promotional offers and the fact that the source allowed the ability to define an offer that will be active in a future time. Unfortunately the protocol and eventually the global schema did not have any concept like that implemented. Furthermore, the global schema belongs to a third party so there was not an option to enhance it with a feature like this. This means that if the source was going to send the relevant data upon offer creation, it would result on false data handling on the global schema. So in case a customer had a promotional offer in the future on their account, a query to display the active offers run on the source and the global schema, would provide inconsistent answers. On the source this offer would still be inactive whereas on the global schema it would show on the results (since it would not have any support for offers in the future).

In other words, by using the theoretical framework we can say that in the integration system $I = \langle G, S, M \rangle$, there are cases where if the data are being sent upon creation, there will be cases where $q_G \nrightarrow q_S$. However, this should not happen for any I, in order to preserve data integrity.

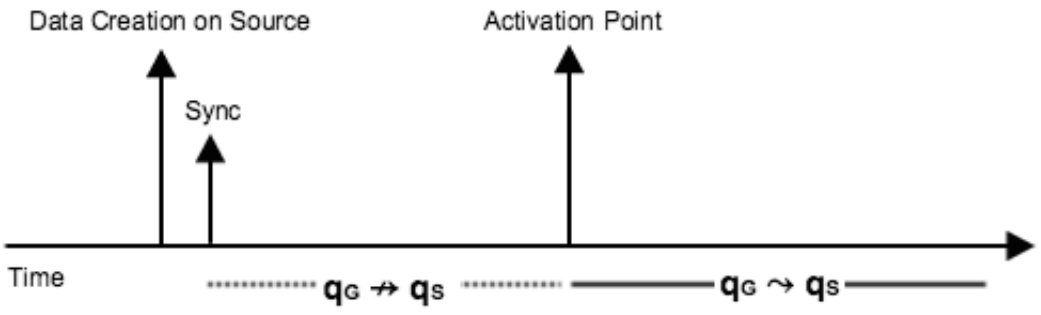

Figure 5.2 The Timing Issue

As figure 5.2 shows, the problem lies on the timing that the data are being synced. After the activation point, the data are properly mapped and there is no issue of misinterpretation from the global schema.

5.5.2 The Solution

To achieve data completeness, we suppressed the synchronisation of those data up to the point that the offer gets enabled in the source. As a result, the data are being synced properly and we have: $q_G \rightsquigarrow q_S$. As figure 5.3 shows, by moving the synchronisation timing to the activation point, the issue is being avoided and we do not have corrupted mappings of data anymore. This happens because the global schema is not aware of the data that exist in the source schema until they reach a state that the global schema itself can interpret them correctly.

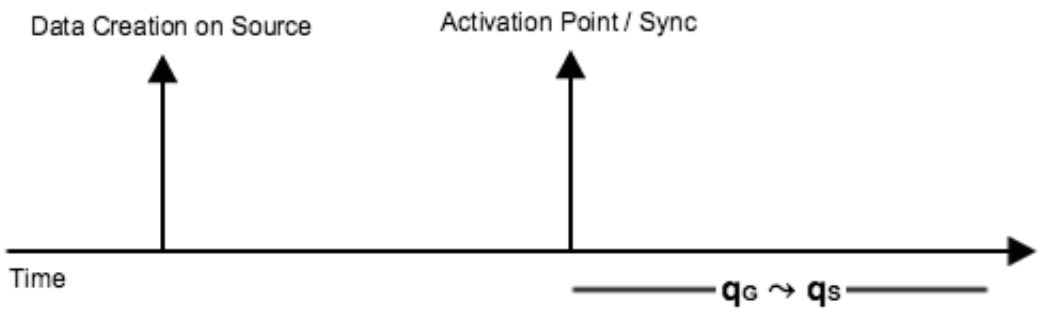

Figure 5.3 Solution for the Timing Issue

5.6 Dependency Issue

The dependency issue arose from the fact that some messages should wait for a master message and according to the result of it, they should either be synced or discarded.

5.6.1 The Problem

Using our case study of the commercial system based on the data warehouse model and protocoled requests we detected a dependency issue which was also related to promotional offers but in a different way. This had to do with the customer's registration and whether it was accepted or not from the integration system. The source could not create promotional offers for a customer that was not synced. But in the real world the company wanted to actually provide offers in the website and then synced upon successful integration of the customer's registration. To make this happen, the offers were being assigned to the customer in a dormant state and on the time that this customer got synced successfully, all the offers were firing up to get awarded and then synced to the global schema.

By using the theoretical framework we can say that in the integration system $I = \langle G, S, M \rangle$, there are cases that if $q_{S_1} \rightsquigarrow q_{G_1}$ is not a successful mapping then $q_{S_2}, q_{S_3} \ldots q_{S_n} \rightsquigarrow q_{G_2}, q_{G_3} \ldots q_{G_n}$ cannot be established until $q_{S_1} \rightsquigarrow q_{G_1}$ is established. Where $q_{S_2}, q_{S_3} \ldots q_{S_n}$ are queries that contain data relevant to q_{S_1} and the same stands for the global schema queries.

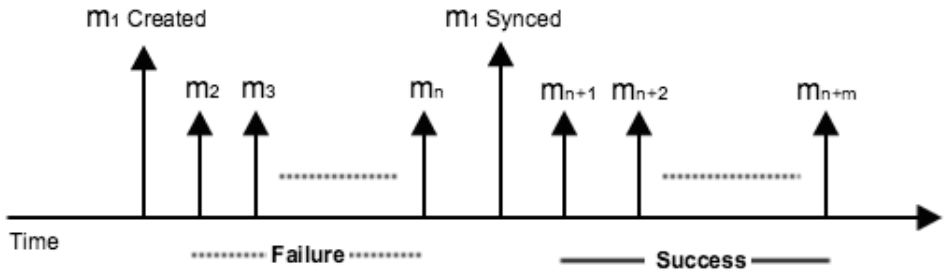

Figure 5.4 The Dependency Issue

In figure 5.4, we demonstrate the situation of the messages before and after the first message (m_1) is sent successfully to the global schema. In this case, $m_2, m_3 \ldots m_n$ are being discarded and $m_{n+1}, m_{n+2} \ldots m_{n+m}$ are being sent successfully. This means that there was no mapping being done for $q_{S_2}, q_{S_3} \ldots q_{S_n} \rightsquigarrow q_{G_2}, q_{G_3} \ldots q_{G_n}$ but there are successful mappings for $q_{S_{n+1}}, q_{S_{n+2}} \ldots q_{S_{n+m}} \rightsquigarrow q_{G_{n+1}}, q_{G_{n+2}} \ldots q_{G_{n+m}}$. As a result there are data in the source schema that never get the chance to be synced over to the global schema.

5.6.2 The Solution

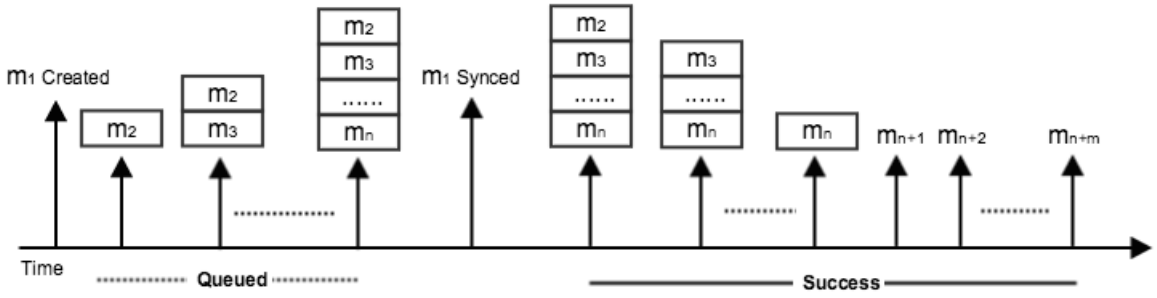

Figure 5.5 Solution for the Dependency Issue

In order to avoid the dependency issue, a solution was implemented which uses a queue of messages that are waiting to be synced. Figure 5.5 describes this solution. As the m_1 is waiting to be synced, the rest of the messages ($m_2, m_3 \ldots m_n$) are being added in a queue. Upon a successful response to m_1, the queue starts processing and eventually all the messages are being sent and the data are being synced successfully. In more detail we can say that this

implementation enables $q_{S_2}, q_{S_3} \ldots q_{S_n} \rightsquigarrow q_{G_2}, q_{G_3} \ldots q_{G_n}$. Also, after the queue is completely processed, upon creation of new data, the synchronisation mechanism keeps working properly and resulting in having $q_{S_{n+1}}, q_{S_{n+2}} \ldots q_{S_{n+m}} \rightsquigarrow q_{G_{n+1}}, q_{G_{n+2}} \ldots q_{G_{n+m}}$ which means that all the data are properly integrated to the global schema throughout the whole process.

In the real world implementation, this solution was applied on user account level. This means that in a single source there can be multiple master messages where each of them is responsible for its own queue that will be processed upon approval. Furthermore, all the queues follow the First In First Out (FIFO) concept which leads to an orderly process of the messages created.

This solution covers the case where m_1 is resulting to a successful response from the global schema. There is also a case where m_1 can fail to receive a successful response. In the scope of this system, this can mean that the user is already registered with another username. In this scenario, the queue is being discarded and the user is being handled as an invalid user for this system. Alternatively, we can say that if m_1 is refused by the global schema, then $m_2, m_3 \ldots m_n$ cannot be synced. Where $m_2, m_3 \ldots m_n$ are dependent to m_1.

5.7 Conclusion

This part of the thesis research addresses two problems on live data integration that can be avoided by implementing solutions that concern the timing and the handling of the integration process. Methods like these can help avoiding the generation of broken mappings between the source and the global schemas. Furthermore they can help on keeping the data consistent between the two schemas and therefore lead to successful integration between the two systems.

While research dealing with similar issues exists, the methodologies suggested are either insufficient to solve the presented problems or they are non-applicable due to the complexity of the systems they study. In the case study of this chapter, simple solutions to the timing

and dependency issues are provided. These solutions can be applied on existing systems that encounter them without the need to resort to modification of the core infrastructure of the data integration system.

From an enterprise point of view, these solutions help generating accurate reports for the status of the data by querying the global schema. The timing issue solution, prevents misinterpretation of the data when querying the global schema. In a similar manner, the solution for the dependency issue results on a complete integration of the data that exist in the source even if initially the global schema fails to sync the data provided.

Since this chapter is dealing with data integration on complex systems, a significant amount of knowledge that regards data integration was gained. The next step was to expand this knowledge even further and apply it to a whole field, namely the field of biological research. The next chapter presents a detailed overview of the current situation of data integration on that field and suggests solutions on how it should ideally be done to improve upon the current situation.

Chapter 6

Data integration in biological research: an overview

6.1 Introduction

Data driven biological research has made data integration strategies crucial for the advancements and discovery in a plethora of fields (e.g. genomics, proteomics, metabolomics, environmental sciences, clinical research to name a few) [79, 135, 166, 167, 181, 200]. Technically, solutions for data integration have been developed and applied in both corporate and academic sectors. When it comes to biological research, there are different interpretations and levels of data integration people seem to consider [46, 47, 76, 101, 103, 130, 146, 196], ranging from genomic data to protein-protein interactions.

Together with data production, there is no doubt that data management, storage and consequently retrieval, analysis and interpretation are at the core of any biological research project. Moreover, the ability to have access to the actual data sets used in a particular study is often crucial for reproducibility and expansion of such study, hence the emphasis in recent years on Open Science and the various initiatives associated [40, 69, 72, 108, 115, 116, 150, 164]. Noticeably, in biological research, the difficulties associated with data integration have

only expanded with the advent of high throughput technologies [111, 135, 168]. Anyone working with Next Generation Sequencing (NGS) faces challenges associated with a variety of aspects this type of data brings, one of the major being: the volume of the data [149, 198].

Here, data integration is referred as the computational solution allowing users, from end user (GUI) to power users (API), to fetch data from different sources, combine, manipulate and re-analyse them as well as being able to create new datasets and share these again with the scientific community.

With this definition in mind, it is clear that data integration solutions are imperative for the advancement of research in biological sciences as well as the mechanisms to make such processes traceable, shareable hence "integrable" [30, 74, 136]. Here, an overview of the strategies most commonly adopted by the biological research community is provided, current challenges and future directions.

6.1.1 Key Concepts and Terminology

Data integration should not just rely on software engineers and computational scientists, but needs to be driven by the actual users whose communities need to define, adopt and use standards, ontologies and annotation best practice. Therefore, it is particularly important for the biological research community to get acquainted with the conceptual basis of data integration, its limitations, challenges and actual terminology.

In order to familiarise the experimental biology community of readers, in Table 6.1.1 key concepts are being presented, definitions and terms used by bioinformaticians and computer scientists.

Schema	A structured and "queryable" way of storing data
Database	A single or collection of schemata

Sources	A number of databases that contain data. Data that reside in each source can either duplicate and/or complement data from other sources
Data Integration	The process of combining data that reside in different sources, to provide users with a unified view of such data
Data Standards	Agreements on representation, format, and definition for common data
Data Formats	A structured way to represent data and metadata in a file
Data Warehousing	Model for integrating data where the data from different sources reside on a central repository (aka data warehouse)
Federated Databases	Model for integrating data where the data reside on the original sources and users are provided with a unified view of the data based on mapping mechanisms of the information
Linked Data	The network of interlinked data that is available on the web. It is used to automatically share semantically rich information and represents the biggest attempt to convert significant amounts of human knowledge across all fields in a computer readable format
Ontology	A structured way of describing data, often presented in a computer-readable format. In bioinformatics, ontologies are sets of unambiguous, universally agreed terms used to describe biological phenomena and "entities", their properties and their relationships
lled Vocabulary	A collection of terms for describing a certain domain of interest

Unique Identifier	A unique representation for a biological entity (molecule, organism, ontology term, etc.). Usually an alphanumeric string that is used to refer to this entity and distinguishes it from others (much like ID or passport number in humans).
Metadata	Data describing data, i.e., additional information (e.g., a comment, explanation, attributes, etc.) for a specific biological entity or process. As an example, in the context of an ontology, this is used to specify significant properties of the ontology
Annotation	The process of attaching relevant information (metadata) to a raw biological entity
Automatic Annotation	Automatic means that the annotation is being done by computer software (often by transferring information from a source to another). This is a way of producing a large amount of metadata
Manual Annotation	As opposed to automatic annotation, manual means that an actual individual does it
GUI	Graphical User Interface. Is the way that a user interacts with a computer by using graphical icons and visual indicators such as buttons, forms etc. In the scope of this chapter the term GUI is being used to refer to interfaces that allow biologists to search/read/edit integrated biological data
API	Application Programming Interface. Set of tool and protocols that a power user can use in order to automatically gain access to functionality and/or data that have been developed/gathered by another individual/organisation

UX	User eXperience. The process of improving user satisfaction by focusing on the usability of a given product.
Visualisation Tools	Applications that help biologists view the data in a more human-friendly way (e.g., Cytoscape for visualising complex networks) like 3D or graph representations of the data

Table 6.1 Terminology

6.2 Review

In computational sciences the theoretical frameworks for data integration have been classified into two major categories namely “eager” and “lazy” [203, 204]. The difference between the two approaches is the way the data get integrated. In the eager approach (warehousing), the data are being copied over to a global schema and stored in a central data warehouse; whereas in the lazy approach the data reside in distributed sources and are integrated on demand based on a global schema used to map the data between sources.

Each of the two main categories of data integration has to deal with its own challenges in order to provide the user with a unified view of the data. In the eager approach, researchers face challenges to keep data updated and consistent, and protect the global schema from having corrupted data [85, 211]. In the lazy approach, data are queried at sources and the scientific community is trying to find ways of improving the answering query process [1, 37, 83, 87, 102, 131] and source completeness [1, 61, 131, 194]. Which approach should be used and when depends on amount of data, who owns them and the existing infrastructure.

In biology a diversity of implementations across these two approaches can be observed that is being used at a variety of levels and forms like data centralisation, federated databases

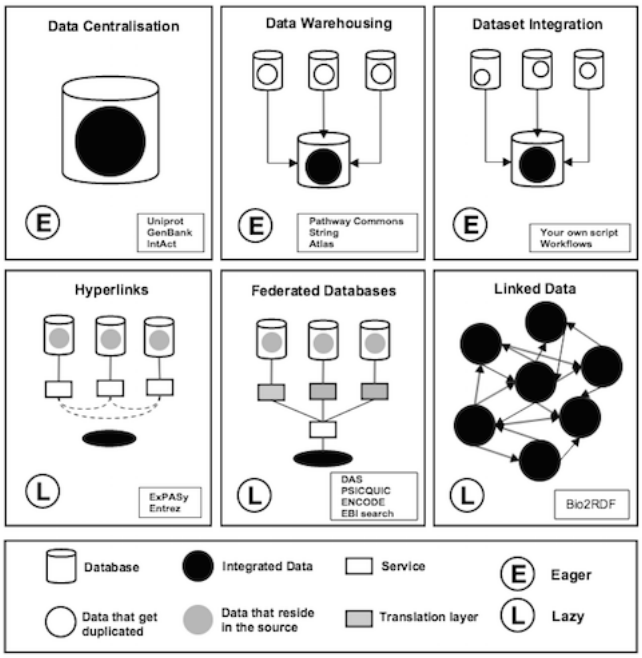

Figure 6.1 **Data integration methodologies.** This figure illustrates six major types of data integration methodologies in biology.

[62, 176] and linked data [20]. Figure 6.1 shows the most common schemata used to integrate data in biology.

UniProt [18] and GenBank [21] are examples of centralised resources (figure6.1-Data Centralisation), whereas Pathway commons [39] collects pathways from different databases and stores them to a shared repository that can be used to query and analyse pathway information (figure 6.1-Data Warehousing). Datasets integration can also be made by in-house workflows accessing distributed databases and downloading data to a local repository (figure 6.1-Dataset Integration). ExPASy [7] is the SIB Bioinformatics Resource Portal through which the user can access databases and tools in different areas of life science (figure 6.1-Hyperlinks). Database links are crucial for interoperability and several efforts have been done in this context [114]. Regarding the federated database model (figure 6.1-Federated Databases), the Distributed Annotation System (DAS) [57] represents a valuable example. DAS is a client-server system used to integrate and display in a single view annotation data

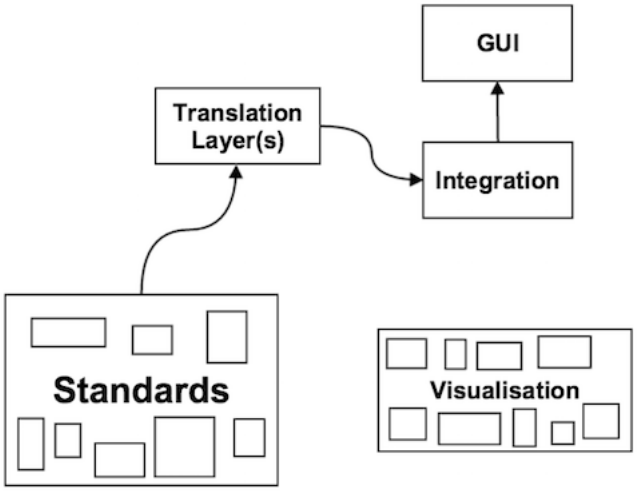

Figure 6.2 **Current state.** This figure illustrates a simplified view of the current state of biological data and tools.

on biological sequences residing over multiple distant servers. In this case, a translation layer is needed to achieve data integration among heterogeneous databases. There are various ways to do this but in general it refers to ways to transform the data from the database to a common format so they can be interpreted in the same way from a mapping service. As for the linked data integration (figure 6.1-Linked Data), the services offered are graphical interfaces (GUI) that provide the user with hyperlinks connecting related data from multiple data providers in a large network of Linked Data. BIO2RDF [20] is an example of such integration system.

Data integration in biological research has its challenges associated to a variety of factors such as standards adoption or easy conversion between data/file formats [79].

Figure 6.2 illustrates a simplified schematic view of the current state of biological research data integration components. Various attempts to integrate the data rely on translation layers that, by applying agreed standards, transform the data in a unified format in order to integrate them. In other words, different formats for the same type of data (e.g. NGS) need to be "translated" into a unified format by applying shared rules. On top of the integration layer, there are various GUIs that make it possible to utilise (download, analyse, represent, etc) the

integrated data. Furthermore, there is a myriad of resources and visualisation tools generated that fail to comply with standards and/or are not compatible with each other [70] On the other hand, controlled vocabularies and ontologies to ease data integration are available for an increasing number of biological domain areas. Some of them can be found at the websites of the OBO (Open Biological and Biomedical Ontologies) foundry [179], the NCBO (National Center for Biomedical Ontology) BioPortal [148], and the OLS (Ontology Lookup Service). One successful example is the XML-based proteomic standards defined by the HUPO-PSI (Human Proteome Organisation-Proteomics Standards Initiative) consortium (see Table 6.2.1). The rest of the chapter will discuss key aspects of standards: ontologies, data formats, identifiers, reporting guidelines, consortiums and standard initiatives which will be followed by a section on visualisation.

6.2.1 Standards

As mentioned above, one of the most important factors for the biological field to thrive is to standardise the data. In computational science a similar problem was encountered for the web and specifically with the way that browsers parse web pages. This was solved by agreeing on W3C standards [24] so that all the browsers are forced to comply otherwise they may result in poor user experience and they risk losing market share.

In biology there are many different ways of representing similar data and this makes the data harder to be integrated and processed to obtain unified views of such data. Gene naming is an example of poor uniformity in data representation. Despite full guidelines were issued in 1979 to adopt gene nomenclature standards (see [99]), an assortment of alternate names is still in use across the scientific literature and databases, posing a challenge to data sharing. When it comes to biological research, it is crucial to create (when non existing), adopt and implement standards. Without these it is (nearly) impossible to achieve data integration [118, 139].

So what exactly are standards? Standards can be defined as an agreed compliant term or structure to represent a biological entity. Entities are all types of units of biological information. For example T, G, A, C is commonly used as a standard way to refer to the nucleotides that make the DNA, and aa (for amino acids) represented usually by one letter, and consequently, a string of letters to represent a DNA or protein sequence. However, a protein might be known in the scientific literature and referred by researchers by a variety of names, synonyms and abbreviations.

So, which standards exist, who defines them and how are these working? Lots of standard initiatives and efforts seem to exist, sometimes redundant, often non driven by the end users communities. It is out of the scope of this chapter (and probably a never ending exercise) to review all of them, which do proliferate but not necessarily in harmonising ways. A snapshot of the variety of standards for metadata can be found at the DCC website [91] and BioSharing [65] as an example of the point that is being made. Table 6.2.1 reports a list of standard initiatives along with their primary goal, URL and key reference in the omics field.

Acronym	**Name**	**Goal**	**URL**	**PMID**
OBO	The Open Biological and Biomedical Ontologies	Establish a set of principles for ontology development to create a suite of orthogonal interoperable reference ontologies in the biomedical domain	http://www.obofoundry.org	17989687

CDISC	Clinical data interchange standards consortium	Establish standards to support the acquisition, exchange, submission and archive of clinical research data and metadata	http://www.cdisc.org	23833735
HUPO-PSI	Human Proteome Organisation - Proteomics Standards Initiative	Defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification	http://www.psidev.info	16901219
GAGH	Global Alliance for Genomics and Health	Create interoperable approaches to catalyze projects that will help unlock the great potential of genomic data	http://genomicsandhealth.org/	24896853
COMBINE	Computational Modeling in Biology	Coordinate the development of the various community standards and formats for computational models	http://co.mbine.org/	25759811

MSI	Metabolomics Standards Initiative	Define community-agreed reporting standards, which provided a clear description of the biological system studied and all components of metabolomics studies	http://msi-workgroups.sourceforge.net	17687353
RDA	Research Data Alliance	Builds the social and technical bridges that enable open sharing of data across multiple scientific disciplines	https://rd-alliance.org	[191]

Table 6.2 List of data standards initiatives

Standards facilitate data re-use. They make data sharing easier, saving overheads and losses of time in data loading, conversion, getting systems to work properly with data. They help overcome interoperability difficulties across different data formats, architectures, and naming conventions, and at infrastructure level, enabling access systems to work together [33, 35, 45, 125, 162]. Absence of standards means substantial loss of productivity and less data available to researchers [43].

Figure 6.3 illustrates a schematic view of an ideal state of biological research data integration components. This figure emphasises on the importance of standards that is the base of all the top layers of the infrastructure. Without solid foundations, it is very difficult to build and maintain robust tools for the layers above. The arrows point out that the data can be used across all layers and this can go both ways. For example, in an ideal state, all biological

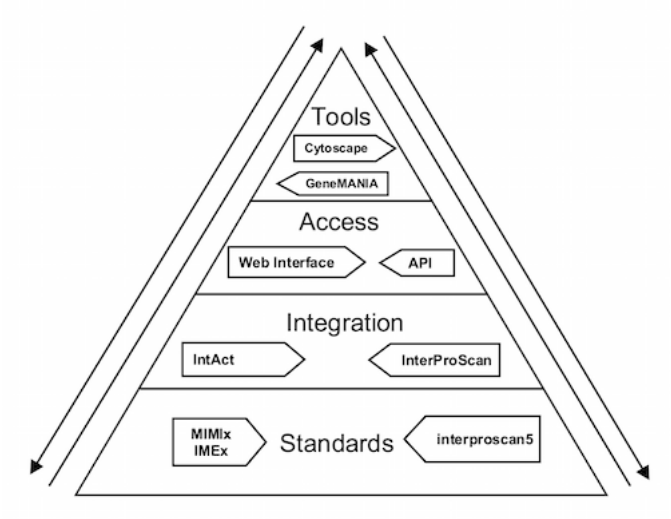

Figure 6.3 **Ideal state.** This figure illustrates a simplified view of an ideal state of biological data and tools.

data would be integrated from various databases across the world and biologists will be able to use a GUI to locate the entity of their interest. Then, they can use a visualisation tool to have a better representation of the entity by using the same data previously identified through the GUI (like a unique identifier). Furthermore, the biologist will be in a position to annotate or edit the data directly from the visualisation tool, which in turn will be able to commit the changes to the integrated service and from then on go all the way down the pyramid until the data in the proper database get edited and annotated.

Standards are therefore key to the data sharing process since they describe the norms which should be adopted to facilitate interchange and inter-working of information, processes, objects and software. Thus data resources play a major role not just in data management, integration, access, and preservation, but also for providing adequate support to research communities [133].

Ontologies

Ontologies have been proliferating in biological research, and their importance underlined several times [14, 42, 140, 178] also in the specific context of data integration [28]. In order

to bring some coordination and consolidation to the proliferation of ontologies across the biological and biomedical research fields, The Open Biological and Biomedical Ontologies (OBO) got together. OBO is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. Biological researchers can get involved and provide feedback by getting into the discussion fora OBO provides. Currently there are ten OBO foundry ontologies and more than 120 candidate ontologies or other ontologies of interest [179].

These efforts need the direct involvement of the actual biologists when it comes to the adoption and implementation of using such ontologies, ensuring these are known and disseminated across communities. Other important initiatives are, the NCBO (National Center for Biomedical Ontology) BioPortal [107, 201], and the OLS (Ontology Lookup Service) [51].

In recent years various ontology matching techniques have been suggested and developed to locate similar ontologies automatically and therefore consider them integrable or even integrate them during the same process [11, 17, 59, 84, 94, 119, 165]. All these methods have high success rates but manual verification is required to consider the data as integrated. This is a good start on trying to normalise biological data.

With a set of unique common compliant standards in place, it will be possible to create tools to integrate the data on the web using an existing infrastructure like linked data. This will enable querying multiple sources without having to re-invent integration techniques for the integration of each source. As an example, one of the efforts currently trying to attempt this is Bio2RDF [20]. This is a major effort to integrate biological data using the linked data infrastructure. So far there are no tools that can utilise these data directly but they are mainly accessible via complex queries or low level GUIs.

6.2.2 Formats

Data formats are the concrete way that humans use to structure and represent biological information in a file. They are particularly relevant to those who deal with large amount of information such that generated by high throughput experiments. Indeed, a scientist interested in a single or a few genes at a time may extract information about them by manually "parsing" the literature or free-text (i.e. non formatted) documents. The need for storing biological data in formatted files arose from the need for using computers to analyse them. The amounts of genomics and proteomics data, which cannot be manually analysed element by element, are exponentially increasing and the adoption of commonly agreed formats to represent them in computer readable files is nowadays of utter importance. Historically, the scarcity of well structured data standards and schemas, caused the flourishing of many different formats even to represent the same type of data despite the adoption of standards in file formats would be essential to data exchange and integration. Funnily, the Roslin Bioinformatics Law's First Law declaims: "The first step in developing a new genetic analysis algorithm is to decide how to make the input data file format different from all pre-existing analysis data file formats" [49].

For the benefit of data integration though, it would be ideal to have well-structured data across few basic formats that would be easily computer readable and therefore easily integrated. In the specific case of NGS data, the lag between the emerging high-throughput screening technologies and the adjusting of the scientific community to settle on a standard format, means time and effort spent on converting raw files across multiple sequencing platforms to make these compatible [13]. Currently, in NGS there are no really "standards" that people adhere to, but a set of commonly used formats (FASTA/Q, SAM, VCF, GFF/GTF, etc.). There are descriptor standards like MIGS [64], but these might not be generally adopted. More in general, today an exhaustive "atlas" of the formats used in bioinformatics cannot be found

Data format class	General data-interchange formats	Nucleotide sequence data	Protein sequence data	Structural data	Sequence alignment	Other data types (PPI, etc)
Table	CSV, TSV	BED; GFF	GFF, Uniprot-GFF	PSF(D), MMCIF(D)	SAM(D)	
FASTA-like		FASTA; FASTQ	FASTA, PIR		SAM(M)	Wig
GenBank-like		GenBank; EMBL	Uniprot-TEXT	PDB, PSF(M), MMCIF(D)	CLUSTAL, MSF, PHYLIP(D)	
Tag-structured	HTML; XML; JSON	SBOL-XML	Uniprot-XML; Uniprot-RDF/XML			PSI MI-XML; PSI-PAR

Table 6.3 Mostly commonly used data formats in bioinformatics. D = data; M = metadata. Formats appearing in more than one class are a mixture of classes.

on the Internet. One partial list is available at http://genome.ucsc.edu/FAQ/FAQformat.html and the description of many formats can be found in the online forum BioStar [158].

A good format needs to take into account the data themselves (for example the DNA sequence of a gene) and the so called metadata, i.e. additional information describing the data (e.g. gene name, taxonomy information, cross reference to other resources, etc.) and has to adopt strategies ("tricks") to make metadata unequivocally distinguishable from data by a computer program. This goal is achieved in different ways by different bioinformatics resources, resulting in the large number of formats that can be observed today. However, despite the large variety of computer readable formats, the most commonly used ones are ascribable to four main different classes: 1) tables 2) FASTA-like 3) GenBank-like 4) tag-structured. Table 6.3 reports examples for each of these classes.

In table formats, data are organised in a table in which the columns are separated by tabs, commas, pipes, etc., depending on the source generating the file. FASTA-like files utilise, for each data record, one or more "definition" or "declaration lines", which contain metadata

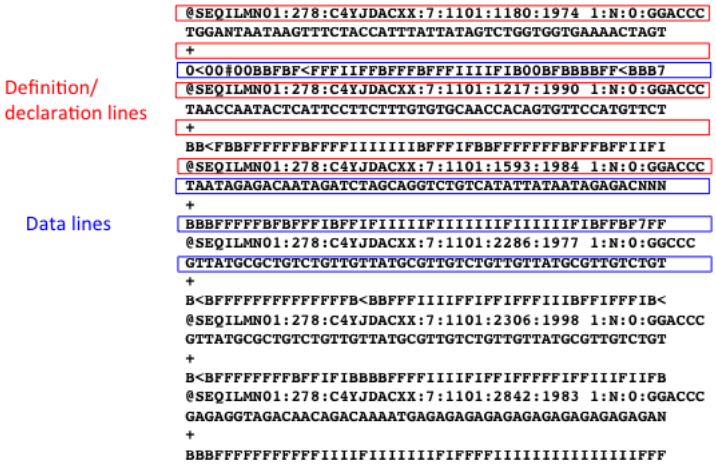

Figure 6.4 **Selected parts of a FASTQ file.** In this format declaration lines start with two different characters ("@" and "+") corresponding to different data types (the raw sequence and the sequence quality values, respectively).

information or specify the content of the following lines. Definition/declaration lines usually start with a special character or keyword in the first position of the line - a “>” in FASTA files or a “@” in fastq or SAM files - followed by lines containing the data themselves (figure 6.4). In some cases, declaration lines may be interspersed with data lines. This format is mostly used for sequence data. In the GenBank-like format, each line starts with an identifier that specifies the content of the line (figure 6.5). Tag-structured formatting uses “tags” (“<”, “>”, “{”, “}”, etc.) to make data and metadata recognisable (figure 6.6) with high specificity. Tag-structured text files, especially XML and JSON, are being increasingly employed as data interchange formats between different programming languages.

There are also examples of data files using different representations for data and metadata. This means that two or more format classes may be used in the same data file. An example

Figure 6.5 **Selected parts of the GenBank entry DQ408531.** The complete entry can be found at http://www.ncbi.nlm.nih.gov/nuccore/DQ408531

Figure 6.6 **Selected parts of the Uniprot entry P01308 in XML format** - The complete entry can be found at http://www.uniprot.org/uniprot/P01308.xml

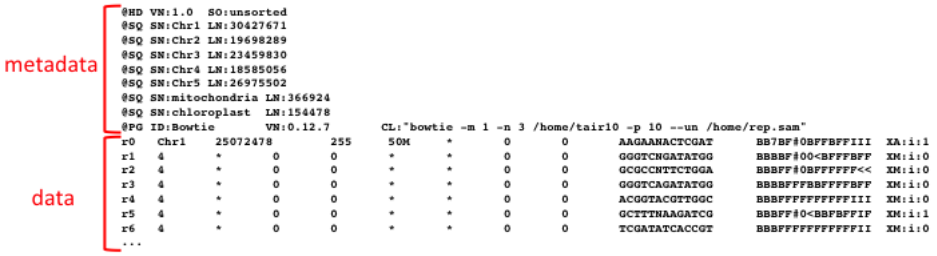

Figure 6.7 Selected parts of a SAM file.

is represented by SAM files, which contain both GenBank-like lines (for the metadata) and table columns (for the data) as shown in figure 6.7.

Should any of these four data representation classes be preferred over the others? Despite the fact that there is an increasing use of XML and some authors propose to adopt XML for biological data interchange between databases and other sources of data [2], there is not an ultimate answer. There are text formats that better suit some specific kind of data and specific computational requirements and purposes. For example, it is difficult to imagine how macromolecule X-ray or NMR coordinates and related annotation, currently stored in PDB files, could fit into the FASTA-like format. On the other hand, if one has to parse big sequence files, the FASTA format, with a single line annotation, will cause them to have a smaller size than differently formatted files and will allow parsing them with just a few lines of code. Notice that some formats (e.g. SAM) can be compressed into a binary version (BAM) for intensive data processing.

Therefore, it is safe to consider that the solution is not to urge scientists to conform to a unique "optimal" format but rather to identify a few operational formats and make database and tool developers aware of the importance of sticking to them.

For integration purposes, the scientific community of database and tool developers has begun to adopt some good practices in data file formatting. One example is represented by the FGED Society (http://fged.org/) formed at a meeting on Microarray Gene Expression Databases (EBI, Hinxton, 1999) with the goal, amongst the others, of facilitating the adoption of standards for DNA microarrays and gene expression data representation. However further efforts should be made in order to achieve a more robust and systematic policy in all the areas where data sharing is essential to utilise these data to make new discoveries and the progress of science possible.

The community of scientists concerned by data sharing and integration, including us, should make the effort of 1) compiling a complete and structured (i.e. organised by data type and purpose) list of the currently available formats with their description and 2) developing guidelines and recommendations for the adoption of standards in file formatting, also discussing which data types fit into each different text format and the related performance implications. This list and the guidelines, which might be integrated in a resource such as BioSharing should encourage database and tool developers to present information in a way that a computer program can parse it, suggest that they avoid inventing new computer readable formats but rather comply with one of the existing ones, and only accept new data, for storage purposes, that meet certain formatting criteria. Such guidelines should be ambitious and forward-looking enough to also advice scientists in both academia and industry to keep in mind data representation in developing high throughput technologies and their information services.

The development of converters translating formats in a unified form should be promoted as well. This would actually make it possible to combine the data across all the formats. A

rather isolated example of data format translation is represented by the PRIDE Converter [15], which makes it easy to translate a large variety of input formats into the unique XML [2, 32] format for proteomic data submission to the PRIDE repository [138]. The PRIDE Converter was designed to be suitable for both small and large data submissions and has a very intuitive GUI also for wet-lab scientists without a strong bioinformatics background or informatics support. Format translation faces problems especially with not well-structured data that cannot be translated properly in a computer readable format and therefore rely on human manipulation of the data in order to verify the correctness of the transformation. In the case of NGS data, people rely on tools for conversion between next generation sequencing data formats, such as NGS-FC (http://sourceforge.net/pro jects/ngsformaterconv/), to ensure each tool in a workflow can work with the right format.

Identifiers

An identifier is a unique representation of a given data entry [22, 113]. For example the Universal Protein Database (UniProt) uses a "unique identifier" to refer to a protein entity which cannot be used in any other case, thus ensuring no redundancy and one agreed unique term that unequivocally identifies a given protein [6].

In biological research a variety of data repositories exist and each of them is using its own implementation for generating unique identifiers. As an example, for the same protein, UniProt uses the identifier Q9Y6N8 whereas Ensembl [54] is referring to it as ENSP00000264463 and RefSeq [163] as NP_006718.2. If all the researchers could use a single unique identifier to refer to a given protein across their publications and work, data integration would be a step ahead of its current state.

An effort to help with the discoverability of the identifiers and assist the researcher with knowledge on how to query data across databases has be done from identifiers.org [109]. This is a registry that facilitates the discovery of resources in life sciences and allows to

decouple the identification of records by the physical locations on the web where they can be retrieved.

Many biological concepts are described in several databases using different identifiers. To facilitate discoverability and integration, databases have their data entries cross-referenced with external entries using identifiers. This enables users to find a data entry like a protein in UniProt and then find the same biological concept described in other databases (ie. RefSeq) and gather more relevant data about the same entry. Several initiatives like PICR [52] or the "DAVID ID conversion tool" [96] provide mapping of such identifiers. It will be beneficial if such service gets integrated in the major bioinformatics databases.

Some organised efforts including distributed resources like IMEx [153] are very well organised and, though the independent databases that are part of the consortium like IntAct [113], MINT [44] and DIP [205] use their own identifiers, all their entries get assigned a unique IMEx identifier issued by a central authority. The IMEx identifier is assigned to a single biological entity with the purpose of being reused across databases/systems and always link to the same entity regardless the system. The IMEx Central repository coordinates curation effort, assigns identifiers and facilitates the exchange of completed records on molecular interaction data between the IMEx Consortium partners.

Approaches like these can increase discoverability and shareability of data and even enable publications and scientific studies to use a single identifier to refer to a given entity. This entity could be easily traced and further studied by their audience. With an infrastructure like this in place, it will be possible to enforce researchers to submit the unique identifier of the biological entity that they are studying on their research papers. This is happening already for nucleotide sequence data where researchers have to submit newly obtained/sequenced entities to one of the three major sequencing databases [128] and refer to it in the paper. Most of other data types can be used in publications without such requirement. This also extends to entire datasets.

Reporting guidelines

Huge steps have been achieved by the creation and adoption of clear recommended guidelines when it comes to depositing and disseminating data and datasets [34, 151, 184, 187]. Such guidelines are often the result of several discussions (years of discussions in some occasions) in a field where data efforts for sharing have been maturing. The specification of several standards in life science include documentation and examples of how to use them, but many initiatives additionally include guidelines to agree on what minimum or recommended information should be provided when describing data. Minimum information guidelines have been very popular to ensure that data can be easily interpreted and that results derived from their analysis can be independently verified. These guidelines tend to concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it. A key landmark in the development of guidelines of minimun information in this area comes from the "Minimum Information about a Biomedical or Biological Investigation" (MIBBI) [187].

It is crucial to have a place where such efforts are listed and shared in order to ensure redundancy is avoided. As an example of reporting guidelines mentioned here the efforts done in the topic of protein-protein interactions. Currently there are two reporting guidelines: MIMIx [154] and IMEx [153]. A key project that is contributing in this area and where one can look for as well as add "reporting guidelines" is the Registry of guidelines in biosharing.org [65, 170].

As mentioned earlier on, there are different formats when it comes to data files, and these will always evolve according to the needs of the communities as well as the nature of the data and associated technologies. For example, a format that contains 20 fields for which one researcher might have a subset of information versus another that might opt for prioritising a different set. It is clear that having a minimum agreed set of fields that all comply to report using standards is crucial for data integration and reusability across such

data. Similarly, other fields might be crucial and informative to a specific set of users. These can be adopted at the level of recommended. For example a protein-protein interaction database wants to capture domain specific information about interactions versus another one that is not interested in such aspect. One also might have optional fields, for those that want to annotate and enrich further the data record with metadata. Doing this in a standard manner means again allowing future reusability and expansion for others to adopt and exchange, integrate data based on this level of information.

Consortiums and standards initiatives

There are several initiatives coordinating the development of community standards to facilitate data comparison, exchange and verification in bioinformatics. Some of this initiatives are community initiatives or consortia like COMBINE [98], PSI [152], GAGH [121], INSDC [147], proteomeXchange [90], IMEx [153], BioPax [55] involved in the development of standards in one specific biological domain. Some other community initiatives like RDA are more generic with a potential application in different scientific domains.

Some strategic efforts supported by major service providers and national governments like ELIXIR [53], BBMRI [208], BD2K [137] are also involved in the development of standards in life sciences. Projects supported by specific grants like BioMedBridges [120], BioSHaRE [56] do also contribute to this cause but their duration is normally bound to the duration of the grant. All these initiatives play a major role in achieving consensus and agreements which facilitates the development and adoptions of standards.

In biological research, molecular biology has been the field ahead in terms of such efforts and the associated bioinformatics applications. One can only imagine the work yet to be done, learning from existing efforts and initiatives as described here in the field of ecology, biodiversity, marine biology and so on. Examples of large scale efforts that need to talk to

each other and ideally apply best practice when it comes to creating an infrastructure that fosters data integration are LifeWatch [16] and ISBE [124].

Visualisation

There is a variety of visualisation tools, but often each tool requires a different file format and the task of feeding back the discovered data is not trivial [71, 160]. The field of visualisation has its own challenges given the increasing quantity of data, the integration of heterogeneous data and the need for tools that allow representing multiple aspects of the data (e.g. multiple connections between nodes with diverse biological meanings [106, 189]). There is a myriad of visualisation and analysis tools, ever proliferating, with each tool providing specific features that address different aspects (e.g. genome browsers [60, 67, 97, 117, 175]). In 2008 Pavlopoulus et al published a wish list for visualisation of biological data which still remains valid [159].

Data integration principles are fundamental in providing tools that are user friendly and allow the end users (biologists) to focus their efforts on the actual study of the data instead of being lost in the process of looking for the data they need by querying multiple databases that appear to provide inconsistent results between them. The field of systems biology *per se* brought substantial advances in visualisations since the ability to analyse and interpret interactions, networks and pathways relies often in the ability of visualising these accurately [159].

Overcoming some of the challenges associated with visualisation relies on better standards adoption and improvement in annotation and metadata. This is clearly a two directional effort: bottom up, where data and datasets are annotated and stored following a common set of standards, this extends to the data formats as well as a top down level of standards and adoption of compatible formats and output files that allow comparisons and integrations of results [3, 93, 182].

Historically, many domains within biology have relied on visualisation as a way to represent the biological information thus creating what are now considered standards in their domains. Plenty of examples can be found in the areas of phylogenetics [161] and pathways [88, 185]. The advent of next generation sequencing brought genomics as a domain were significant effort has been put to develop new visualisation techniques to represent sequences, alignments, expression patterns and ultimately entire genomes [81, 95, 143, 199]. However, biological researchers might lack an understanding and awareness about the range of visualisation techniques available and which is the most appropriate visual representation [183, 197].

An increased dialogue between the computational scientists involved in the creation and development of such tools with the end users (aka the biologists), would be beneficial for the entire community and hopefully this work is one step towards such outcome. Efforts in this direction are also on the way and it is worth citing here the BiVi initiative (http://bivi.co/), which is addressing several challenges in the realm of visualisation as well as trying to reduce the gap between the biology, computational sciences and developers of bioinformatics tools. BiVi has grouped many of the most notable visualisation tools produced by biologists and developers across seven domains (though some of the tools cover more than one of these) and provides information as to their provenance, current status and links to websites (http://bivi.co/visualisations). Other community efforts in this area are VizBI (http://vizbi.org/), SciVis (http://scivis.itn.liu.se/) and CoVis (http://www.iwr.uni-heidelberg.de/groups/CoVis/).

It would be impossible for us to list the plethora of visualisation tools developed and used in biological research, hence an overview is provided in Table6.4 of some of the most common visualisations tools in the area of “Interaction Network Visualisation” to illustrate the variety and types of resources available for one area.

Name of resource	What it does	URL
BicOverlapper	Visualisation of biclusters combined with profile plots and heat maps	http://vis.usal.es/bicoverlapper/
BiGGEsTS	Heat map-based bicluster visualisation	http://tinyurl.com/BiGGEsTS
Brain Explorer	Visualisation of 3D transcription data in the central nervous system	http://tinyurl.com/brainExplorer
Data Matrix Viewer	Simple profile plot visualisation; supports Gaggle	http://gaggle.systemsbiology.net/
EXPANDER	Heat maps, scatter plots and profile plots of cluster averages	http://acgt.cs.tau.ac.il/expander
GENESIS	Analysis suite; offers several interactive visualisations	http://genome.tugraz.at/
geWorkbench	Modular suite; heat maps, dendrograms, profile and scatter plots	http://tinyurl.com/geWorkbench
Hierarchical Clustering Explorer	Linked heat map, profile and scatter plots; systematic exploration	http://tinyurl.com/HCExplorer
Java TreeView	Linked heat maps, karyoscopes, sequence alignments, scatter plots	http://jtreeview.sourceforge.net/
Mayday	Modular suite; many linked visualisations; enhanced heat map113	http://tinyurl.com/maydaywp
MultiExperiment Viewer	Analysis suite; heat maps, dendrograms, profile and scatter plots	http://www.tm4.org/
PointCloudXplore	Visualisation of 3D transcription data in Drosophila embryos	http://tinyurl.com/PointCloudXplore
TimeSearcher	Exploration and analysis of time series; advanced profile plots	http://tinyurl.com/timesearcher
R/BioConductor Geneplotter	Karyoscope-style plots and other visualisations	http://www.bioconductor.org/
GenePattern	Modular analysis platform; several visualisation modules available	http://tinyurl.com/GenePatt
Cytoscape	Open source software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data	http://www.cytoscape.org/index.html

Table 6.4 Common visualisations tools in the area of "Interaction Network Visualisation"

There are also well known and generally adopted analysis suites that also provide visualisation tools as part of their repertoire of resources such as Galaxy [73], Cytoscape [174, 180], Ondex [122], iPlant Collaborative [75], Bioconductor [134]. Other important efforts derive from initiatives that are working towards unlocking the actual visualisations, in other words going from the visualisation to the data and datasets. This is important not only for reproducibility but also to allow access for data and their integration with other data/datasets. A very interesting resource is Utopia Docs [8, 9], a free PDF reader that connects the static content of scientific articles to the dynamic world of online content. This resources allows the user to interact directly with curated database entries; play with molecular structures; edit sequence and alignment data; even plot and export tabular data. Another totally different but relevant initiative in the world of visualisation is BIOJS, that aims to provide open-source library of JavaScript components to visualise biological data. BIOJS vision is that every online biological dataset in the world should be visualised with BIOJS tools (http://biojs.net/) [78].

6.3 Conclusion

6.3.1 Data integration strategies on biological research

When it comes to data integration, there are many ways of implementing a system architecture in order to provide the user with a unified view of the data. In the field of biological research it seems that they adopted all of the possible ways of data integration. More specifically biologists are using both the eager and lazy approaches and all of the techniques to achieve them as well. This shows from one aspect a huge effort of the biological community to integrate data but on the other hand it resulted to great data heterogeneity.

6.3.2 Current challenges

With the current state of the biological data, it seems that data heterogeneity is one of the biggest challenges in biological data integration. There are too many standard initiatives that refer to the same biological entities. This overcomplicates the attempts to integrate data. Furthermore usually the standards are being generated without properly structured research groups to provide knowledge from each of the fields involved. This is why with the exception of some specific sub-fields of biological research (like genomics), the data between researchers and institutes present a big heterogeneity.

6.3.3 Suggestions and future directions

In order to overcome the challenges mentioned above, there should be a change in how the source and integration system are being analysed and developed. Biologists should get more involved with the aspects described here and working with bioinformaticians and computational scientists to achieve uniformity of their data. With this issue resolved, integration of biological data will greatly boost biological research and the field will gain a more robust structure: computational scientists will be responsible for maintaining and improving the infrastructure of the data; bioinformaticians will be able to build upon this infrastructure; biologists will be able to do research with advanced tools without the overhead of getting acquainted with complex topics of database management and programming tools.

6.3.4 What is next

By researching the current state of biological data, it became clear that one important aspect that can help the biologists and bioinformaticians to get their data in an ideal state is training. To contribute on this aspect, an open source application was created that simplifies the discovery of training materials for bioinformatics. The next chapter presents a detailed overview of this effort.

Chapter 7

BATMat: Bioinformatics Autodiscovery of Training Materials

7.1 Introduction

Getting a clear overview of bioinformatics training materials available online is not a trivial task [105, 172]. Existing efforts such as the Global Organisation for Bioinformatics Learning, Education & Training (GOBLET) [10, 50] and the ELIXIR Training e-Support System (http://elixir-uk.org/training-platform) allow training portal and cross-reference indexing for bioinformatics training materials. These resources may be the source of training materials or may provide pointers to the source they originated from. However, most users still rely on the Google search engine for retrieval go materials with keywords of interest. Currently this process can be time consuming and daunting, risking considerable unproductive browsing and clicking before finding useful content. BATMat is an open source query tool for training materials based on Google's custom search API specifically tailored to bioinformatics. Using a precompiled list of relevant web resources where bioinformatics training content is known to be included (e.g., GOBLET, Coursera, etc.). BATMat can easily be incorporated into existing third party training portals to provide users with the option to extend the search

beyond the items that are already included in their repositories. Therefore it harmonizes discoverability and shareability of resources from which such materials originate. BATMat can serve as a central point that collects data on the fly from relevant sites and provides a snapshot of the training available at a particular time. Its use is expected to be relevant to trainers, students, funders and regulators.

BATMat can be easily incorporated in existing efforts and provide users with the option to look for materials that have been manually uploaded as well as obtaining an overview of further items on the web related to their query subject automatically. This effort will also push for best practice in annotation of materials, since it is clear from the current situation that, the lack of harmonisation leads to a lack of discoverability, shareability and ultimately a loss in terms of the resources used to create such items in the first instance. It is foreseeable that more variables can be included as the community needs evolve. For example, if all the materials uploaded on the various websites would use a common terminology to indicate the audience they were design to, it would be possible to search and sort or even filter them according to their target audiences. This is something that could be easily implemented in BATMat, however the heterogeneity of materials and lack of concerted effort is what makes this currently impracticable.

7.2 Bioinformatics Autodiscovery of Training Materials

BATMat provides an integrated, coherent solution that gives a snapshot in an organised manner of the available training materials in a specific subject (via a query term). BATMat only displays minimal information, such as title, description, source URL and file(s) associated. BATMat's philosophy is that all open source available training materials should be discoverable and listed only once. The community of bioinformatics, trainers and students who use it will shape BATMat's evolution and functionalities. At the moment, BATMat provides (see Figure 7.1):

- A customised search using Google's "custom search API" specifically focused on bioinformatics training websites. This is used to retrieve the URL, title and description of the results.
- Dual display options: tabular or Google like search.
- Results that can be sorted according to metadata such as: website URL (source) and website title.
- Retrieved hits can be filtered according to their content (i.e., whether they contain associated files or not).
- A registry of known bioinformatics training resources on which the search is performed, together with a controlled interface for their editing and addition of new resources.
- Suite of web services and widgets that can be embedded in external websites, to display pre-programmed filtered lists of hits automatically.
- Third party request for adding new websites to be included in the query.

In figure 7.1 a usage example is being presented. The term "genomics" is typed in the query search box (1) and a search performed by selecting either the "Search" or "Files Only" button (2). This selection will determine the type of output BATMat will provide. In this case "Search" is hit and both web content and files are retrieved. The type of output can be displayed in a table or Google-like layout (3). The corresponding search results (4) provide information regarding the site, title of the material or relevant website, its description and files (if any).

Technically, the application is written in PHP and it is powered by the Google Custom Search API. In the backend it forwards the query to the API and gets the results. When this happens, it shows the results to the user and then it starts parsing each one of them to look for training materials which later are being sent back to the user asynchronously. Figure 7.2

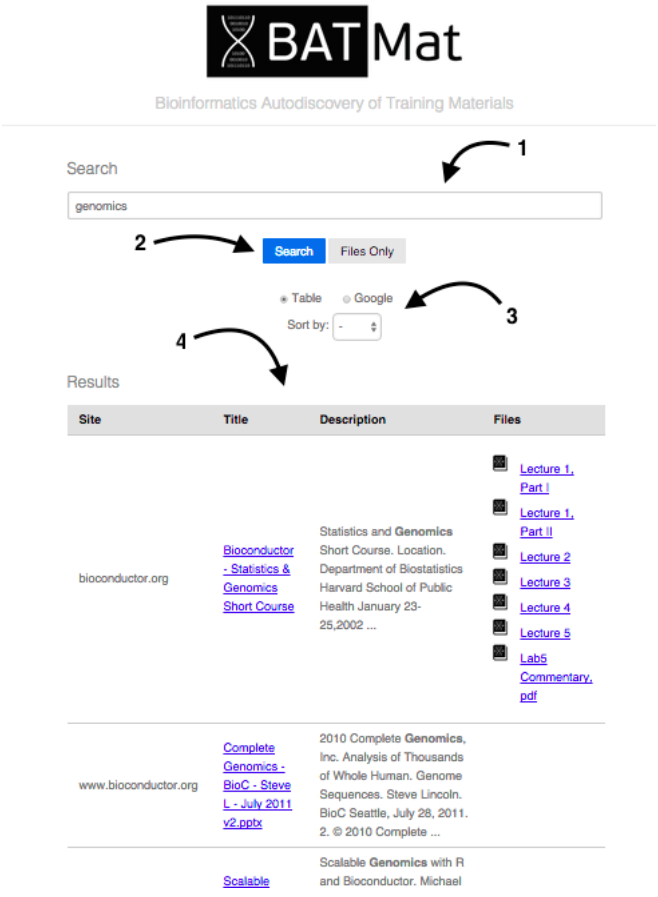

Figure 7.1 BATMat Usage example.

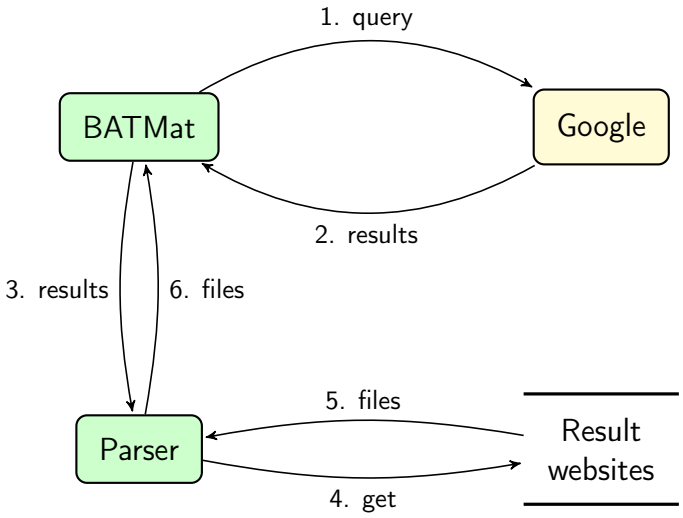

Figure 7.2 BATMat's data flow diagram

is the data flow diagram of the application. Also, in order to achieve better service in case the web page parsing takes a long time, a caching layer was created that stores all the Google requests and queries in a JSON file on the server side. On every request, this file is being checked before an actual search or website request is being attempted.

7.3 The Importance of Community Driven Standards & Ontologies

There is an increasing trend for bioinformatics training materials to be shared online as a way to gain credit, and reciprocal sharing of materials that can be reutilised under a lenient license. Except for a few exceptions such as Coursera, GOBLET and TeSS, sharing of materials is often done without any strategic dissemination plan, leading to great heterogeneity in which materials are described, distributed and made accessible, even within the same institution. This is an unfortunate situation that leads to reduced discoverability, wasting considerable time and leads to loss of attribution of credit where it is due. As a solution

to this problem, BATMat provides a core set of functionalities that encourages standards adoption and, ultimately, shareability of bioinformatics materials in a completely accessible manner. BATMat's capabilities markedly contrast with the current poor discoverability of existing materials via ordinary Google searches.

7.4 Conclusion

More than ever there is an increase in training needs in bioinformatics and data analysis skills related to data-driven science. The ability for anyone to get with a click an organised table with recent materials is more than timely. BATMat is a bioinformatics auto-discovery search engine for training materials that is freely accessible. BATMat can be accessed directly at http://imbatmat.com and all the code is deposited in a Bit Bucket repository (https://bitbucket.org/Pr301/batmat). BATMat's code (Apache License, version 2.0) is open source can be reused and easily customised. Its evolution and impact will be determined by the community that adopts it.

Chapter 8

Conclusion

People's daily lives are increasingly immersed into the way they work, communicate and interact with the world through the web. Research done in the time window of this thesis already reflects the speed and advancements of the web in terms of technologies and applications. The overall conclusion is that integration technologies are vital to the advancement of web applications. This fact is more pronounced in the advent of high-throughput content generation and data mining. This final part summarizes the topics that were covered by this thesis in the order they appeared in previous chapters and reflects on their outcomes and conclusions.

Chapter 2 investigated ways to utilise desktop data like users' bookmarks in order to generate tailored user profiles. It proposed an application that would figure out the user's interests based on the content of the bookmarks by means of integration with on-line classification web services. This is something which has been proven successful and it has been applied on a major web browser, Google Chrome. The results of the evaluation of the application showed that a decent level of accuracy can be achieved without complex implementation. At the same time, the study explored the privacy issues that arise due to fact that the user data have to cross the boundaries between local computer and public internet.

Chapter 3 explored and proposed a conceptual set of advanced web technologies to enhance the experience of the learners and tutor in an e-Learning environment. In particular, it proposed a series of interactive tools and resources that would allow synchronous contributions from learners and tutors but also asynchronous recovery of information. In other words, a record of the interactivity is constructed so that the learner can re-experience -or observe if it was missed- the actual interactive session. Some of the concepts introduced in this chapter have been applied in Content Learning Environments such as Moodle, Drupal and Blackboard. However, some features mentioned here are yet to be seen in real case scenarios. In particular, the "Wiki-enabled Interface" and the "Customised background search" would be interesting to be seen in use in the education (undergrads) and training (postgrads) realms, considering the increasing use and adoption of Wikipedia entries as source of information by the current and future generations. Also it will be interesting to see powerful search tools utilized in automated ways in e-learning applications.

Chapter 4 explored how combining desktop data with web sources from media social networking services can bring powerful and trustworthy features to desktop computing. Features like these are present in various applications especially when they are connected to an on-line store (iTunes, Spotify etc). But the concept of integrating web features on the actual desktop search engine is something that has been very recently adopted by Apple's Spotlight. Similarly, Microsoft is using this concept with Cortana. Both applications are newer compared to the first time this concept was proposed herein in 2011.

Chapter 5 addressed two problems on live data integration and showed how these problems can be avoided by simple manipulation of the timing that the integration process occurs. The outcome of this work provides an example on how such approaches can aid avoiding the generation of broken mappings between a source and a global schema.

Chapter 6 illustrated more than ever the importance of data integration and the significant role that advanced web applications play in the field of biological research. This chapter

highlights the bottleneck and challenges faced in this area, illustrating that it is not the technical aspects or the absence of advanced web technologies that hinder integration but rather the actual luck of understanding and adoption of web standards by the biological community. In other words, these are fields where standards and ontologies are not yet mature and diverse contributors to web applications have no formal informatics background. On the contrary, they are often domain experts with self taught programming skills, risking to invest in developments that are not maintainable or usable in the long term.

Chapter 7 illustrated how by customizing a web query and targeting the search to specific more relevant sites to the academic field of bioinformatics, the user receives an optimized and more efficient search results. Such customization approaches are key to the data driven era of science, including bioinformatics training. Increasing the effectiveness of the user's web experience leads to overall gains in terms of resources, efficiency and actual accuracy of the retrieved information.

Concluding, it is evident nowadays that the understanding and ability to develop the web further in terms of technologies, applications and data integration requires an increasing involvement of the computer scientists across disciplines. Useful web applications, integrating data from diverse web sources, are key parts to the future advances of the web. The research goals of the present work pursued the discovery and validation of methodologies for the implementation of such cutting edge web applications. The outcomes and concepts originated from this thesis research on data integration are considered successful, since similar approaches were later adopted and applied by major and dominant informatics development companies in the world.

Future work on the subject will focus on experimenting with new methodologies and technologies of data integration in modern web applications. With the pace that web technologies evolve, novel research will be continuously needed in order to achieve effective data integration across disciplines and between complex scientific and commercial computer

systems and applications. The author plans to pursue the aforementioned goals by extending the line of research of the current thesis towards intelligent ways of data integration.

Bibliography

[1] Abiteboul, S. and Duschka, O. M. (1998). Complexity of answering queries using materialized views. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 254–263. ACM.

[2] Achard, F., Vaysseix, G., and Barillot, E. (2001). Xml, bioinformatics and data integration. *Bioinformatics*, 17(2):115–125.

[3] Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., Casas, E., Cheng, H. H., Clarke, L., Couldrey, C., Dalrymple, B. P., Elsik, C. G., Foissac, S., Giuffra, E., Groenen, M. A., Hayes, B. J., Huang, L. S., Khatib, H., Kijas, J. W., Kim, H., Lunney, J. K., McCarthy, F. M., McEwan, J. C., Moore, S., Nanduri, B., Notredame, C., Palti, Y., Plastow, G. S., Reecy, J. M., Rohrer, G. A., Sarropoulou, E., Schmidt, C. J., Silverstein, J., Tellam, R. L., Tixier-Boichard, M., Tosser-Klopp, G., Tuggle, C. K., Vilkki, J., White, S. N., Zhao, S., and Zhou, H. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.*, 16:57.

[4] Apple Inc. (2011a). itunes - http://www.apple.com/itunes.

[5] Apple Inc. (2011b). Spotlight - http://www.apple.com/macosx/what-is-macosx/spotlight.html.

[6] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115–D119.

[7] Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., and Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.*, 40(Web Server issue):597–603.

[8] Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., and Thorne, D. (2009). Calling International Rescue: knowledge lost in literature and data landslide! *Biochem. J.*, 424(3):317–333.

[9] Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., and Thorne, D. (2010). Utopia documents: linking scholarly literature with research data. *Bioinformatics*, 26(18):i568–574.

[10] Atwood, T. K., Bongcam-Rudloff, E., Brazas, M. E., Corpas, M., Gaudet, P., Lewitter, F., Mulder, N., Palagi, P. M., Schneider, M. V., van Gelder, C. W., et al. (2015). Goblet: The global organisation for bioinformatics learning, education and training.

[11] Aumueller, D., Do, H.-H., Massmann, S., and Rahm, E. (2005). Schema and ontology matching with coma++. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 906–908. ACM.

[12] Aumüller, D. and Auer, S. (2005). Towards a semantic wiki experience-desktop integration and interactivity in wiksar. In *Semantic Desktop Workshop*, pages 2005–2012.

[13] Baker, M. (2010). Next-generation sequencing: adjusting to data overload. *nature methods*, 7(7):495–499.

[14] Bard, J. B. and Rhee, S. Y. (2004). Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*, 5(3):213–222.

[15] Barsnes, H., Vizcaino, J. A., Eidhammer, I., and Martens, L. (2009). Pride converter: making proteomics data-sharing easy. *Nature biotechnology*, 27(7):598–599.

[16] Basset, A. and Los, W. (2012). Biodiversity e-science: Lifewatch, the european infrastructure on biodiversity and ecosystem research. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology*, 146(4):780–782.

[17] Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi, M. (2008). Bgee: integrating and comparing heterogeneous transcriptome data among species. In *Data Integration in the Life Sciences*, pages 124–131. Springer.

[18] Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-Uhalte, E., Silva, A. D., De Giorgi, M., Dogan, T., Fazzini, F., Gane, P., Castro, L. G., Garmiri, P., Hatton-Ellis, E., Hieta, R., Huntley, R., Legge, D., Liu, W., Luo, J., MacDougall, A., Mutowo, P., Nightingale, A., Orchard, S., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zellner, H., Cowley, A., Figueira, L., Li, W., McWilliam, H., Lopez, R., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., de Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Suzek, B. E., Vinayaka, C., Wang, Q., Wang, Y., Yeh, L. S., Yerramalla, M. S., and Zhang, J. (2015). UniProt: a hub for protein information. *Nucleic Acids Res.*, 43(Database issue):D204–212.

[19] Baykan, E., Henzinger, M., Marian, L., and Weber, I. (2009). Purely url-based topic classification. In *Proceedings of the 18th international conference on World wide web*, pages 1109–1110. ACM.

[20] Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform*, 41(5):706–716.

[21] Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2015). GenBank. *Nucleic Acids Res.*, 43(Database issue):D30–35.

[22] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2000). GenBank. *Nucleic Acids Res.*, 28(1):15–18.

[23] Benz, D., Tso, K., and Schmidt-Thieme, L. (2006). Automatic bookmark classification: A collaborative approach. In *Proceedings of the Second Workshop on Innovations in Web Infrastructure (IWI 2006), Edinburgh, Scotland*, volume 76.

[24] Berjon, R., Faulkner, S., Leithead, T., Pfeiffer, S., O'Connor, E., and Navara, E. D. (2014). HTML5. Candidate recommendation, W3C. http://www.w3.org/TR/2014/CR-html5-20140731/.

[25] Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., and Weitzner, D. J. (2006). A framework for web science. *Foundations and trends in Web Science*, 1(1):1–130.

[26] Bizer, C. (2009). The emerging web of linked data. *Intelligent Systems, IEEE*, 24(5):87–92.

[27] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far.

[28] Blake, J. A. and Bult, C. J. (2006). Beyond the data deluge: data integration and bio-ontologies. *J Biomed Inform*, 39(3):314–320.

[29] Box, D., Ehnebuske, D., Kakivaya, G., Layman, A., Mendelsohn, N., Nielsen, H. F., Thatte, S., and Winer, D. (2000). Simple object access protocol (soap) 1.1.

[30] Bravo, E., Calzolari, A., De Castro, P., Mabile, L., Napolitani, F., Rossi, A. M., and Cambon-Thomsen, A. (2015). Developing a guideline to standardize the citation of bioresources in journal articles (cobra). *BMC medicine*, 13(1):33.

[31] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (1998). Extensible markup language (xml). *World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210*, page 16.

[32] Bray, T., Sperberg-McQueen, M., Paoli, J., Yergeau, F., and Maler, E. (2004). Extensible markup language (XML) 1.0 (third edition). W3C recommendation, W3C. http://www.w3.org/TR/2004/REC-xml-20040204.

[33] Brazma, A. (2001). On the importance of standardisation in life sciences. *Bioinformatics*, 17(2):113–114.

[34] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, 29(4):365–371.

[35] Brooksbank, C. and Quackenbush, J. (2006). Data standards: a call to action. *OMICS*, 10(2):94–99.

[36] Brusilovsky, P. and Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education*, 13(2):159–172.

[37] Calvanese, D., De Giacomo, G., Lenzerini, M., and Vardi, M. Y. (2000). Answering regular path queries using views. In *Data Engineering, 2000. Proceedings. 16th International Conference on*, pages 389–398. IEEE.

[38] Canonical Ltd. (2011). Integrated desktop search - https://wiki.ubuntu.com/integrateddesktopsearch.

[39] Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, 39(Database issue):D685–690.

[40] Chamberlain, S. A. and Szocs, E. (2013). taxize: taxonomic search and retrieval in R. *F1000Res*, 2:191.

[41] Chan, J. O. (2015). Optimizing data warehousing startegies. *Communications of the IIMA*, 5(1):1.

[42] Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent systems*, 14(1):20–26.

[43] Charalabidis, Y., Gonçalves, R. J., and Popplewell, K. (2010). Developing a science base for enterprise interoperability. In *Enterprise Interoperability*, volume IV, pages 245–254. Springer.

[44] Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, 35(Database issue):D572–574.

[45] Chervitz, S. A., Deutsch, E. W., Field, D., Parkinson, H., Quackenbush, J., Rocca-Serra, P., Sansone, S. A., Stoeckert, C. J., Taylor, C. F., Taylor, R., and Ball, C. A. (2011). Data standards for Omics data: the basis of data sharing and reuse. *Methods Mol. Biol.*, 719:31–69.

[46] Cheung, K.-H., Yip, K. Y., Smith, A., Masiar, A., Gerstein, M., et al. (2005). Yeasthub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics*, 21(suppl 1):i85–i96.

[47] Chung, S. Y. and Wong, L. (1999). Kleisli: a new tool for data integration in biology. *Trends in Biotechnology*, 17(9):351–355.

[48] Cormode, G. and Krishnamurthy, B. (2008). Key differences between web 1.0 and web 2.0. *First Monday*, 13(6).

[49] Corpas, M., Fatumo, S., and Schneider, R. (2012). How not to be a bioinformatician. *Source Code Biol Med*, 7(1):3.

[50] Corpas, M., Jimenez, R. C., Bongcam-Rudloff, E., Budd, A., Brazas, M. D., Fernandes, P. L., Gaeta, B., van Gelder, C., Korpelainen, E., Lewitter, F., et al. (2014). The goblet training portal: a global repository of bioinformatics training materials, courses and trainers. *Bioinformatics*, page btu601.

[51] Cote, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J. A., and Hermjakob, H. (2010). The Ontology Lookup Service: bigger and better. *Nucleic Acids Res.*, 38(Web Server issue):W155–160.

[52] Cote, R. G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., and Hermjakob, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, 8:401.

[53] Crosswell, L. C. and Thornton, J. M. (2012). ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.*, 30(5):241–242.

[54] Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kahari, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R., and Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Res.*, 43(Database issue):D662–669.

[55] Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, O., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Ruebenacker, O., Reubenacker, O., Samwald, M., van Iersel, M., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K. H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Le Novere, N., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., and Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, 28(9):935–942.

[56] Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H., Perola, M., Stolk, R. P., Foco, L., Minelli, C., Waldenberger, M., Holle, R., Kval?y, K., Hillege, H. L., Tasse,

A. M., Ferretti, V., and Fortier, I. (2013). Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg Themes Epidemiol*, 10(1):12.

[57] Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R., and Stein, L. (2001). The distributed annotation system. *BMC Bioinformatics*, 2:7.

[58] Dunn, M., Gwertzman, J., Layman, A., and Partovi, H. (1997). Privacy and Profiling on the Web. Note, W3C. http://www.w3.org/TR/NOTE-Web-privacy.html.

[59] Ehrig, M. and Staab, S. (2004). Qom–quick ontology mapping. In *The Semantic Web–ISWC 2004*, pages 683–697. Springer.

[60] Engels, R., Yu, T., Burge, C., Mesirov, J. P., DeCaprio, D., and Galagan, J. E. (2006). Combo: a whole genome comparative browser. *Bioinformatics*, 22(14):1782–1783.

[61] Etzioni, O., Golden, K., and Weld, D. S. (1997). Sound and efficient closed-world reasoning for planning. *Artificial Intelligence*, 89(1–2):113 – 148.

[62] Etzold, T. and Argos, P. (1993). SRS–an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, 9(1):49–57.

[63] Fehler, S. and Bilodeau, R. (2007). Web 2.0 principles and best practices.

[64] Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glockner, F. O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S. E., Li, K., Lister, A. L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S. A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San Gil, I., Wilson, G., and Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.*, 26(5):541–547.

[65] Field, D., Sansone, S., Delong, E. F., Sterk, P., Friedberg, I., Gaudet, P., Lewis, S., Kottmann, R., Hirschman, L., Garrity, G., Cochrane, G., Wooley, J., Meyer, F., Hunter, S., White, O., Bramlett, B., Gregurick, S., Lapp, H., Orchard, S., Rocca-Serra, P., Ruttenberg, A., Shah, N., Taylor, C., and Thessen, A. (2010). Meeting Report: BioSharing at ISMB 2010. *Stand Genomic Sci*, 3(3):254–258.

[66] Franz, T., Dellschaft, K., and Staab, S. (2009). *Unlock Your Data: The Case of MyTag*. Springer.

[67] Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, 32(Web Server issue):W273–279.

[68] Friedman, M., Levy, A., and Millstein, T. (1999). Navigational plans for data integration. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 67–73, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[69] Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P. E., and Gil, Y. (2013). Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. *PLoS ONE*, 8(11):e80278.

[70] Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., et al. (2010a). Visualization of omics data for systems biology. *Nature methods*, 7:S56–S68.

[71] Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., and Gavin, A. C. (2010b). Visualization of omics data for systems biology. *Nat. Methods*, 7(3 Suppl):56–68.

[72] Gentleman, R. and Temple Lang, D. (2004). Statistical analyses and reproducible research.

[73] Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, 15(10):1451–1455.

[74] Goble, C. and Stevens, R. (2008). State of the nation in data integration for bioinformatics. *J Biomed Inform*, 41(5):687–693.

[75] Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., Lowry, S., Mock, S., Helmke, M., Kubach, A., Narro, M., Hopkins, N., Micklos, D., Hilgert, U., Gonzales, M., Jordan, C., Skidmore, E., Dooley, R., Cazes, J., McLay, R., Lu, Z., Pasternak, S., Koesterke, L., Piel, W. H., Grene, R., Noutsos, C., Gendler, K., Feng, X., Tang, C., Lent, M., Kim, S. J., Kvilekval, K., Manjunath, B. S., Tannen, V., Stamatakis, A., Sanderson, M., Welch, S. M., Cranston, K. A., Soltis, P., Soltis, D., O'Meara, B., Ane, C., Brutnell, T., Kleibenstein, D. J., White, J. W., Leebens-Mack, J., Donoghue, M. J., Spalding, E. P., Vision, T. J., Myers, C. R., Lowenthal, D., Enquist, B. J., Boyle, B., Akoglu, A., Andrews, G., Ram, S., Ware, D., Stein, L., and Stanzione, D. (2011). The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front Plant Sci*, 2:34.

[76] Goldovsky, L., Janssen, P., Ahren, D., Audit, B., Cases, I., Darzentas, N., Enright, A. J., Lopez-Bigas, N., Peregrin-Alvarez, J. M., Smith, M., Tsoka, S., Kunin, V., and Ouzounis, C. A. (2005). CoGenT++: an extensive and extensible data environment for computational genomics. *Bioinformatics*, 21(19):3806–3810.

[77] Golfarelli, M., Rizzi, S., and Cella, I. (2004). Beyond data warehousing: What's next in business intelligence? In *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP*, DOLAP '04, pages 1–6, New York, NY, USA. ACM.

[78] Gomez, J., Garcia, L. J., Salazar, G. A., Villaveces, J., Gore, S., Garcia, A., Martin, M. J., Launay, G., Alcantara, R., Del-Toro, N., Dumousseau, M., Orchard, S., Velankar, S., Hermjakob, H., Zong, C., Ping, P., Corpas, M., and Jimenez, R. C. (2013). BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics*, 29(8):1103–1104.

[79] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8(Suppl 2):I1.

[80] Google Inc. (2011). Google desktop - http://desktop.google.com/features.html.

[81] Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.*, 36(10):3420–3435.

[82] Grahne, G. and Mendelzon, A. O. (1999a). Tableau techniques for querying information sources through global schemas. In b1, editor, *Database Theory ICDT'99*, volume b2 of *b4*, b5 b6, pages 332–347. Springer, b7, b8 edition.

[83] Grahne, G. and Mendelzon, A. O. (1999b). Tableau techniques for querying information sources through global schemas. In *In Proc. of the 7th Int. Conf. on Database Theory (ICDT'99)*, volume 1540 of *Lecture Notes in Computer Science*, pages 332–347, Springer.

[84] Gross, A., Hartung, M., Kirsten, T., and Rahm, E. (2010). On matching large life science ontologies in parallel. In *Data Integration in the Life Sciences*, pages 35–49. Springer.

[85] Gupta, A. and Widom, J. (1993). Local verification of global integrity constraints in distributed databases. In *ACM SIGMOD International Conference on Management of Data (SIGMOD 1993)*.

[86] Halevy, A., Rajaraman, A., and Ordille, J. (2006). Data integration: The teenage years. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, pages 9–16. VLDB Endowment.

[87] Halevy, A. Y. (2001). Answering queries using views: A survey. *The VLDB Journal*, 10(4):270–294.

[88] Haw, R., Hermjakob, H., D'Eustachio, P., and Stein, L. (2011). Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics*, 11(18):3598–3613.

[89] Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.

[90] Hermjakob, H. and Apweiler, R. (2006). The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev Proteomics*, 3(1):1–3.

[91] Higgins, S. (2008). The dcc curation lifecycle model. *International Journal of Digital Curation*, 3(1):134–140.

[92] Hirsch, C., Hosking, J., Grundy, J., Chaffe, T., MacDonald, D., and Halytskyy, Y. (2009). The visual wiki: A new metaphor for knowledge access and management. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–10. IEEE.

[93] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., and Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature*, 455(7209):47–50.

[94] Hu, W., Qu, Y., and Cheng, G. (2008). Matching large ontologies: A divide-and-conquer approach. *Data & Knowledge Engineering*, 67(1):140–160.

[95] Huang, d. a. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57.

[96] Huang, d. a. W., Sherman, B. T., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2008). DAVID gene ID conversion tool. *Bioinformation*, 2(10):428–430.

[97] Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The ensembl genome database project. *Nucleic acids research*, 30(1):38–41.

[98] Hucka, M., Nickerson, D. P., Bader, G. D., Bergmann, F. T., Cooper, J., Demir, E., Garny, A., Golebiewski, M., Myers, C. J., Schreiber, F., Waltemath, D., and Le Novere, N. (2015). Promoting Coordinated Development of Community-Based Information Standards for Modeling in Biology: The COMBINE Initiative. *Front Bioeng Biotechnol*, 3:19.

[99] HUGO Gene Nomenclature Committee (2015). - http://www.genenames.org/about/overview.

[100] Hull, R. (1997). Managing semantic heterogeneity in databases: a theoretical prospective. In *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 51–61. ACM.

[101] Hwang, D., Rust, A. G., Ramsey, S., Smith, J. J., Leslie, D. M., Weston, A. D., De Atauri, P., Aitchison, J. D., Hood, L., Siegel, A. F., et al. (2005). A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17296–17301.

[102] Ives, Z. G., Florescu, D., Friedman, M., Levy, A., and Weld, D. S. (1999). An adaptive query execution system for data integration. In *ACM SIGMOD Record*, volume 28, pages 299–310. ACM.

[103] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453.

[104] Jiang, X. and Tan, A.-H. (2005). Mining ontological knowledge from domain-specific text documents. In *Data Mining, Fifth IEEE International Conference on*, pages 4–pp. IEEE.

[105] Jimenez, R. C., Albar, J. P., Bhak, J., Blatter, M.-C., Blicher, T., Brazas, M. D., Brooksbank, C., Budd, A., De Las Rivas, J., Dreyer, J., et al. (2013). iann: an event sharing platform for the life sciences. *Bioinformatics*, 29(15):1919–1921.

[106] Johnson, C., Moorhead, R., Munzner, T., Pfister, H., Rheingans, P., and Yoo, T. S. (2006). Nih/nsf visualization research challenges report.

[107] Jonquet, C., Lependu, P., Falconer, S., Coulet, A., Noy, N. F., Musen, M. A., and Shah, N. H. (2011). NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. *Web Semant*, 9(3):316–324.

[108] Juty, N., Ali, R., Glont, M., Keating, S., Rodriguez, N., Swat, M., Wimalaratne, S., Hermjakob, H., Le Novère, N., Laibe, C., et al. (2015). Biomodels: Content, features, functionality, and use. *CPT: Pharmacometrics & Systems Pharmacology*, 4(2):1–14.

[109] Juty, N., Le Novere, N., and Laibe, C. (2012). Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.*, 40(Database issue):D580–586.

[110] Kabassi, K. and Virvou, M. (2003). Using web services for personalised web-based learning. *Educational Technology & Society*, 6(3):61–71.

[111] Kadadi, A., Agrawal, R., Nyamful, C., and Atiq, R. (2014). Challenges of data integration and interoperability in big data. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 38–40. IEEE.

[112] Kan, M.-Y. and Thi, H. O. N. (2005). Fast webpage classification using url features. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 325–326. ACM.

[113] Karp, P. D. (1996a). A protocol for maintaining multidatabase referential integrity. *Pac Symp Biocomput*, pages 438–445.

[114] Karp, P. D. (1996b). Database links are a foundation for interoperability. *Trends Biotechnol.*, 14(8):273–279.

[115] Kauppinen, T. and de Espindola, G. M. (2011). Linked open science-communicating, sharing and evaluating data, methods and results for executable papers. *Procedia Computer Science*, 4:726–731.

[116] Kenall, A., Edmunds, S., Goodman, L., Bal, L., Flintoft, L., Shanahan, D. R., and Shipley, T. (2015). Better reporting for better research: a checklist for reproducibility. *BMC neuroscience*, 16(1):44.

[117] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, 12(6):996–1006.

[118] Kher, S., Dickerson, J., and Rawat, N. (2010). Biological pathway data integration trends, techniques, issues and challenges: A survey. In *Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on*, pages 177–182. IEEE.

[119] Kirsten, T., Thor, A., and Rahm, E. (2007). Instance-based matching of large life science ontologies. In *Data Integration in the Life Sciences*, pages 172–187. Springer.

[120] Klech, H., Brooksbank, C., Price, S., Verpillat, P., Buhler, F. R., Dubois, D., Haider, N., Johnson, C., Linden, H. H., Payton, T., Renn, O., and See, W. (2012). European initiative towards quality standards in education and training for discovery, development and use of medicines. *Eur J Pharm Sci*, 45(5):515–520.

[121] Knoppers, B. M. (2014). International ethics harmonization and the global alliance for genomics and health. *Genome Med*, 6(2):13.

[122] Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P., and Philippi, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390.

[123] Kolaitis, P. G. (2005). Schema mappings, data exchange, and metadata management. In *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 61–75, New York, NY, USA. ACM.

[124] Krajewski, P., Chen, D., Ćwiek, H., van Dijk, A. D., Fiorani, F., Kersey, P., Klukas, C., Lange, M., Markiewicz, A., Nap, J. P., et al. (2015). Towards recommendations for metadata and data handling in plant phenotyping. *Journal of experimental botany*, page erv271.

[125] Lacroix, Z. (2003). *Bioinformatics: managing scientific data*, volume 6. Academic Press.

[126] Lapatas, V. and Stefanidakis, M. (2010). Combining desktop data and web 3.0 technologies to profile a user. In *WEBIST (1)*, pages 350–353.

[127] Lawton, G. (2008). Moving the os to the web. *IEEE Computer*, 41(3):16–19.

[128] Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., et al. (2010). The european nucleotide archive. *Nucleic acids research*, page gkq967.

[129] Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA. ACM.

[130] Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004). Smart 4.0: towards genomic data integration. *Nucleic Acids Research*, 32(suppl 1):D142–D144.

[131] Levy, A. Y. (1996). Obtaining complete answers from incomplete databases. In *VLDB*, volume 96, pages 402–412. Citeseer.

[132] Levy, A. Y. (2000). Logic-based techniques in data integration. In *Logic-based artificial intelligence*, pages 575–595. Springer.

[133] Lord, P. et al. (2006). Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models, the joint data standards study, http://www. mrc. ac. uk/pdfjdss_final_report. pdf.

[134] Luo, W. and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831.

[135] Ma'ayan, A., Rouillard, A. D., Clark, N. R., Wang, Z., Duan, Q., and Kou, Y. (2014). Lean big data integration in systems biology and systems pharmacology. *Trends in pharmacological sciences*, 35(9):450–460.

[136] Mabile, L., Dalgleish, R., Thorisson, G. A., Deschênes, M., Hewitt, R., Carpenter, J., Bravo, E., Filocamo, M., Gourraud, P. A., Harris, J. R., et al. (2013). Quantifying the use of bioresources for promoting their sharing in scientific research. *GigaScience*, 2(1):1–8.

[137] Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., and Green, E. D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*, 21(6):957–958.

[138] Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005). Pride: the proteomics identifications database. *Proteomics*, 5(13):3537–3545.

[139] Mathew, J. P., Taylor, B. S., Bader, G. D., Pyarajan, S., Antoniotti, M., Chinnaiyan, A. M., Sander, C., Burakoff, S. J., and Mishra, B. (2007). From bytes to bedside: Data integration and computational biology for translational cancer research. *PLoS computational biology*, 3(2):e12.

[140] Mayer, G., Jones, A. R., Binz, P.-A., Deutsch, E. W., Orchard, S., Montecchi-Palazzi, L., Vizcaíno, J. A., Hermjakob, H., Oveillero, D., Julian, R., et al. (2014). Controlled vocabularies and ontologies in proteomics: overview, principles and practice. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1844(1):98–107.

[141] McAfee, A. P. (2006). Enterprise 2.0: The dawn of emergent collaboration. *MIT Sloan management review*, 47(3):21–28.

[142] Microsoft Corp. (2011). Windows search - http://www.microsoft.com/windows/products/winfamily/desktopsearch/default.mspx.

[143] Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F., and Trajanoski, Z. (2005). PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.*, 33(Web Server issue):W633–637.

[144] Möller, K. and Decker, S. (2005). Harvesting desktop data for semantic blogging.

[145] Murray-Rust, P., Adams, S. E., Downing, J., Townsend, J., and Zhang, Y. (2011). The semantic architecture of the world-wide molecular matrix (wwmm). *J. Cheminformatics*, 3:42.

[146] Myers, C. L. and Troyanskaya, O. G. (2007). Context-sensitive data integration and prediction of biological networks. *Bioinformatics*, 23(17):2322–2330.

[147] Nakamura, Y., Cochrane, G., and Karsch-Mizrachi, I. (2013). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, 41(Database issue):D21–24.

[148] NCBO BioPortal (2015). - http://bioportal.bioontology.org/.

[149] Nekrutenko, A. and Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672.

[150] Neylon, C. and Wu, S. (2009). Open science: tools, approaches, and implications. In *Pacific symposium on biocomputing*, volume 14, pages 540–544. Citeseer.

[151] Orchard, S., Al-Lazikani, B., Bryant, S., Clark, D., Calder, E., Dix, I., Engkvist, O., Forster, M., Gaulton, A., Gilson, M., Glen, R., Grigorov, M., Hammond-Kosack, K., Harland, L., Hopkins, A., Larminie, C., Lynch, N., Mann, R. K., Murray-Rust, P., Lo Piparo, E., Southan, C., Steinbeck, C., Wishart, D., Hermjakob, H., Overington, J., and Thornton, J. (2011). Minimum information about a bioactive entity (MIABE). *Nat Rev Drug Discov*, 10(9):661–669.

[152] Orchard, S., Hermjakob, H., and Apweiler, R. (2003). The proteomics standards initiative. *Proteomics*, 3(7):1374–1376.

[153] Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F. S., Brinkman, F., Cesareni, G., Chatr-aryamontri, A., Chautard, E., Chen, C., Dumousseau, M., Goll, J., Hancock, R. E., Hancock, R., Hannick, L. I., Jurisica, I., Khadake, J., Lynn, D. J., Mahadevan, U., Perfetto, L., Raghunath, A., Ricard-Blum, S., Roechert, B., Salwinski, L., Stumpflen, V., Tyers, M., Uetz, P., Xenarios, I., and Hermjakob, H. (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, 9(4):345–350.

[154] Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stumpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., Salama, J. J., Moore, S., Wojcik, J., Bader, G. D., Vidal, M., Cusick, M. E., Gerstein, M., Gavin, A. C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J., Prieto, C., Perreau, V. M., Hogue, C., Mewes, H. W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G., and Hermjakob, H. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, 25(8):894–898.

[155] O'reilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1):17.

[156] Pahl, C. and Holohan, E. (2009). Applications of semantic web technology to support learning content development. *Interdisciplinary Journal of E-Learning and Learning Objects*, 5.

[157] PAN, T., ZHENG, L.-n., ZHANG, H.-j., FANG, C.-b., LOU, J., and SHAO, Z. (2009). Combining web services toward innovative design of agile virtual enterprise supported by web 3.0. *WSEAS Transactions on Communications*, 8(1):81–91.

[158] Parnell, L. D., Lindenbaum, P., Shameer, K., Dall'Olio, G. M., Swan, D. C., Jensen, L. J., Cockell, S. J., Pedersen, B. S., Mangan, M. E., Miller, C. A., and Albert, I. (2011). BioStar: an online question & answer resource for the bioinformatics community. *PLoS Comput. Biol.*, 7(10):e1002216.

[159] Pavlopoulos, G. A., Wegener, A. L., and Schneider, R. (2008). A survey of visualization tools for biological network analysis. *BioData Min*, 1:12.

[160] Pettifer, S., Thorne, D., McDermott, P., Marsh, J., Villeger, A., Kell, D. B., and Attwood, T. K. (2009). Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinformatics*, 10 Suppl 6:S19.

[161] Phylogeny Programs (2015). - http://evolution.genetics.washington.edu/phylip/software.html.

[162] Piwowar, H. A., Becich, M. J., Bilofsky, H., and Crowley, R. S. (2008). Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Med.*, 5(9):e183.

[163] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 35(Database issue):D61–65.

[164] Ram, K. (2013). ropensci-open tools for open science. In *AGU Fall Meeting Abstracts*, volume 1, page 04.

[165] Rance, B., Gibrat, J.-F., and Froidevaux, C. (2009). An adaptive combination of matchers: application to the mapping of biological ontologies for genome annotation. In *Data Integration in the Life Sciences*, pages 113–126. Springer.

[166] Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T. E., and Nilges, M. (2007). ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, 23(3):381–382.

[167] Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97.

[168] Saleem, M., Kamdar, M. R., Iqbal, A., Sampath, S., Deus, H. F., and Ngomo, A.-C. N. (2014). Big linked cancer data: Integrating linked tcga and pubmed. *Web Semantics: Science, Services and Agents on the World Wide Web*, 27:34–41.

[169] Sampson, D. G. (2009). Competence-related metadata for educational resources that support lifelong competence development programmes. *Educational Technology & Society*, 12(4):149–159.

[170] Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., et al. (2012). Toward interoperable bioscience data. *Nature genetics*, 44(2):121–126.

[171] Sauermann, L. (2005). The gnowsis semantic desktop for information integration. In *Wissensmanagement*, pages 39–42. Citeseer.

[172] Schneider, M. V., Watson, J., Attwood, T., Rother, K., Budd, A., McDowall, J., Via, A., Fernandes, P., Nyronen, T., Blicher, T., et al. (2010). Bioinformatics training: a review of challenges, actions and support requirements. *Briefings in bioinformatics*, 11(6):544–551.

[173] Sen, A. and Sinha, A. P. (2005). A comparison of data warehousing methodologies. *Commun. ACM*, 48(3):79–84.

[174] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504.

[175] Shannon, P. T., Reiss, D. J., Bonneau, R., and Baliga, N. S. (2006). The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics*, 7:176.

[176] Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). Biomart–biological queries made easy. *BMC genomics*, 10(1):22.

[177] Smith, A. K., Cheung, K.-H., Yip, K. Y., Schultz, M., and Gerstein, M. B. (2007a). Linkhub: a semantic web system that facilitates cross-database queries and information retrieval in proteomics. *BMC bioinformatics*, 8(Suppl 3):S5.

[178] Smith, B. (2003). The logic of biological classification and the foundations of biomedical ontology. In *Invited Papers from the 10th International Conference in Logic Methodology and Philosophy of Science, Oviedo, Spain*, pages 19–25.

[179] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007b). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, 25(11):1251–1255.

[180] Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

[181] Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012). An encyclopedia of mouse dna elements (mouse encode). *Genome biology*, 13(8):418.

[182] Stein, L. (2001). Genome annotation: from sequence to biology. *Nat. Rev. Genet.*, 2(7):493–503.

[183] Stobbe, M. D., Jansen, G. A., Moerland, P. D., and van Kampen, A. H. (2014). Knowledge representation in metabolic pathway databases. *Brief. Bioinformatics*, 15(3):455–470.

[184] Sweet, J. J. (2014). Editorial. EQUATOR - reporting guidelines for "Enhancing the QUality and Transparency Of health Research". *Clin Neuropsychol*, 28(4):547–548.

[185] Tanabe, M. and Kanehisa, M. (2012). Using the KEGG database resource. *Curr Protoc Bioinformatics*, Chapter 1:Unit1.12.

[186] Tao, X., Li, Y., Zhong, N., and Nayak, R. (2007). Ontology mining for personalized web information gathering. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 351–358. IEEE.

[187] Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P.-A., Bogue, M., Booth, T., et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the mibbi project. *Nature biotechnology*, 26(8):889–896.

[188] Teevan, J., Dumais, S. T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456. ACM.

[189] Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics*, 14(2):178–192.

[190] Thuraisingham, B. (2002). Data mining, national security, privacy and civil liberties. *ACM SIGKDD Explorations Newsletter*, 4(2):1–5.

[191] Treloar, A. (2014). The research data alliance: Globally co-ordinated action against barriers to data publishing and sharing. *Learned Publishing*, 27(5):9–13.

[192] Ullman, J. D. (1997). Information integration using logical views. In *Database Theory—ICDT'97*, pages 19–40. Springer.

[193] URL-Classifier (2009). Url classification service - http:///www.urlclassifier.com.

[194] van der Meyden, R. (1998). Logical approaches to incomplete information: A survey. In Chomicki, J. and Saake, G., editors, *Logics for Databases and Information Systems*, pages 307–356. Kluwer.

[195] Vlamos, P., Floros, A., Giannakos, M. N., and Drossos, K. C. (2010). Towards an interactive e-learning system based on emotions and affective cognition. In *proceedings of International Conference on Information Communication Technologies in Education, ICICTE 2010*.

[196] Von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B., and Bork, P. (2007). String 7—recent developments in the integration and prediction of protein interactions. *Nucleic acids research*, 35(suppl 1):D358–D362.

[197] Walter, T., Shattuck, D. W., Baldock, R., Bastin, M. E., Carpenter, A. E., Duce, S., Ellenberg, J., Fraser, A., Hamilton, N., Pieper, S., Ragan, M. A., Schneider, J. E., Tomancak, P., and Heriche, J. K. (2010). Visualization of image data from cells to organisms. *Nat. Methods*, 7(3 Suppl):26–41.

[198] Wandelt, S., Rheinländer, A., Bux, M., Thalheim, L., Haldemann, B., and Leser, U. (2012). Data management challenges in next generation sequencing. *Datenbank-Spektrum*, 12(3):161–171.

[199] Wang, J., Zhang, Y., Marian, C., and Ressom, H. W. (2012). Identification of aberrant pathways and network activities from high-throughput data. *Brief. Bioinformatics*, 13(4):406–419.

[200] Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., and Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, 38(Web Server issue):W214–220.

[201] Whetzel, P. L. (2013). NCBO Technology: Powering semantically aware applications. *J Biomed Semantics*, 4 Suppl 1:S8.

[202] Widom, J. (1995a). Research problems in data warehousing. In *Proceedings of the Fourth International Conference on Information and Knowledge Management*, CIKM '95, pages 25–30, New York, NY, USA. ACM.

[203] Widom, J. (1995b). Research problems in data warehousing. In *Proceedings of the fourth international conference on Information and knowledge management*, pages 25–30. ACM.

[204] Widom, J. (1996). Integrating heterogeneous databases: Lazy or eager? *ACM Comput. Surv.*, 28(4es).

[205] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, 30(1):303–305.

[206] Yao, Y. Y. (1995). Measuring retrieval effectiveness based on user preference of documents. *J. Am. Soc. Inf. Sci*.

[207] Yu, L. and Liu, Y. (2015). Using linked data in a heterogeneous sensor web: challenges, experiments and lessons learned. *International Journal of Digital Earth*, 8(1):15–35.

[208] Yuille, M., van Ommen, G. J., Brechot, C., Cambon-Thomsen, A., Dagher, G., Landegren, U., Litton, J. E., Pasterk, M., Peltonen, L., Taussig, M., Wichmann, H. E., and Zatloukal, K. (2008). Biobanking for Europe. *Brief. Bioinformatics*, 9(1):14–24.

[209] Zepeda, J. S. and Chapa, S. V. (2007). From desktop applications towards ajax web applications. In *Electrical and Electronics Engineering, 2007. ICEEE 2007. 4th International Conference on*, pages 193–196. IEEE.

[210] Zhou, B. and Yao, Y. (2010). Evaluating information retrieval system performance based on user preference. *Journal of Intelligent Information Systems*, 34(3):227–248.

[211] Zhuge, Y., García-Molina, H., Hammer, J., and Widom, J. (1995). View maintenance in a warehousing environment. *SIGMOD Rec.*, 24(2):316–327.