Journal of Biomedical Informatics 41 (2008) 687–693

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

State of the nation in data integration for bioinformatics

Carole Goble *, Robert Stevens

School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK

ARTICLE INFO

Article history:
Received 22 October 2007
Available online 5 February 2008

Keywords:
Databases
Data integration
Data warehouse
Life sciences
Link
Mashup
RDF
Semantics
View integration
Workflow

ABSTRACT

Data integration is a perennial issue in bioinformatics, with many systems being developed and many technologies offered as a panacea for its resolution. The fact that it is still a problem indicates a persistence of underlying issues. Progress has been made, but we should ask "what lessons have been learnt?", and "what still needs to be done?" Semantic Web and Web 2.0 technologies are the latest to find traction within bioinformatics data integration. Now we can ask whether the Semantic Web, mashups, or their combination, have the potential to help.

This paper is based on the opening invited talk by Carole Goble given at the Health Care and Life Sciences Data Integration for the Semantic Web Workshop collocated with WWW2007. The paper expands on that talk. We attempt to place some perspective on past efforts, highlight the reasons for success and failure, and indicate some pointers to the future.

1. Introduction

Bioinformatics is a discipline based on a wealth of diverse, complex and distributed data resources. It is not a surprise that data integration has been discussed as a major challenge in bioinformatics for many years, ever since the beginning of dataset collection and distributed publication. It seems that as a discipline, bioinformatics is rather proud of the growing number of resources it holds (nearly 900 reported in [1]). On the contrary, perhaps we should be a little ashamed. Stein argues that we need to build a "Bioinformatics Nation" from the competing, fractured "princely states" of the current situation [2]. Moreover, the *integration* of resources—a prerequisite for most bioinformatics analysis—is a perennial and costly challenge. The presence of Data Integration in the Life Sciences (DILS),[1] a thriving annual forum dedicated to publishing yet more new integration systems, attests to the importance of the issue.

To gain some kind of perspective on data integration in bioinformatics, we need to examine what it is in the nature of bioinformatics that perpetuates this integration challenge. We need to reflect upon the progress that has been made, identify the kinds of integration regimes that have been tried and understand how they have met this challenge. We can also speculate upon how the new technologies of the Semantic Web and Web 2.0 "mashups" might help (or hinder) integration in bioinformatics.

2. A loose federation of bio-nations

Integration is necessary due to the large, and increasing, numbers of data resources within bioinformatics. The annual *Nucleic Acids Research* journal database supplement listed 96 databases in 2001 [3] and 800+ in 2007 [1]. A fundamental question for bioinformatics database integration is "why so many data resources?" Here are some reasons: There is an ecosystem of primary data collections feeding secondary and tertiary databases. The Web makes it (too) easy to publish. Being a resource provider is one way to make a reputation. There are many types of data and each has its own communities and its own repositories. Each new sub-discipline that arises develops its own data representations skewed to its own biases. The distributed and diverse nature of the discipline promotes a "long tail" of specialist resource suppliers rather than the few centralised data centres we see in disciplines such as particle physics. Consequently, each type of data has a multiplicity of resources, many replicating, partially overlapping or presenting slightly different views on more or less the same data types. For example, there are some 231 different pathway databases[2]; different types of data will necessitate more than one pathway resource, but this number seems excessive. It appears that it is easier, more

* Corresponding author. Fax: +44 161 275 6235.
E-mail addresses: carole.goble@manchester.ac.uk (C. Goble), robert.stevens@manchester.ac.uk (R. Stevens).

[1] http://dils07.cis.upenn.edu/.

[2] http://www.pathguide.org.

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.01.008

desirable, or more expedient, to create a database afresh than it is adapt or re-use existing resources.

The biology research groups spawning these databases are highly autonomous. When a group or individual decides to make a new data resource, sometimes without the technical or modelling skills of a database designer, a different data model and different values are frequently given to the data. In addition, the delivery mechanism will vary: flat-files (of arbitrary "ASCII art"); a multitude of XML schema; and highly variable Application Programming Interfaces (APIs). Ignorance—intentional or otherwise—of the customers of the resources often leads to the disruptive churn of interfaces, schemas and formats. Again, this is a symptom of the decoupled nature of the discipline; those that use the resources are frequently independent and decoupled from those who create them. Thus, the autonomous nature of much bioinformatics data management, together with the volatile nature of the data and the fast moving nature of experimental developments within the science (mass sequencing, transcriptomics, proteomics, all in the past decade) means that there is a tendency for multiple, highly heterogeneous data resources to appear. The volatility of the databases themselves (Merali and Giles [4] report that only 18% of databases surveyed had a sustained future) means that many also disappear to be replaced by others providing the same or similar resources, usually in a different fashion.

3. Making a hard problem even harder

Bioinformatics touts its quantity of data as a problem. However, compared to other disciplines such as particle physics or astronomy, biological data are modestly sized. Rather, the important distinctive feature is the *complexity* of those data (despite that complexity sometimes being self-inflicted). Complexity arises from: describing a sample and its originating context; the processing of a sample in an experiment; the diversity of sources for a sample; the variability of data quality and evidence/trust levels; the diversity in types of data; the large number of interlinked collections of these types; the changeable nature of the data.

The drive for integration has to tackle this complexity. The following headings cover topics that have to be addressed.

- *The need for common, shared identities and names.* No matter what integration scheme or method is favoured all approaches need to match up records referring to the same data object. For example, the WS-1 protein has ten different names and 21 distinct accession numbers.[3] As Mike Ashburner famously said, "a biologist would rather share their toothbrush than their gene name." Although Stein in [5] proposed the notion of *de facto* naming authorities with distinct namespaces, this is still a source of confusion and difficulty. Two different database entries are clearly different, but whether the entity or entities they represent are identical or equivalent is challenging. For example, a Uniprot [6] entry refers to a class of proteins (via a representation of its primary structure), a class of variant proteins or some viral protein, whereas a KEGG [7] entry refers to a collection of proteins involved in some pathway. To link these, a bioinformatician must map a Uniprot entry's protein (sequence) to a KEGG entry. The syntax and mechanics of a single global unique identity scheme, and who should be a naming authority, is still a subject of vigorous debate. There are continuing arguments over the relative merits of identity technologies, including the Life Science identifier [8], Persistent URLs[4] or some other scheme. This uncertainty is hindering progress.
- *The need for shared semantics.* Coordinating resources that have differing conceptualisations and representations makes bioinformatics data analysis harder than it need be. For example, a pseudogene is a gene-like structure containing in-frame stop codons, or is a transposable cassette that is rearranged, or includes a full open reading frame but is not transcribed. The community recognises the need for community standards for both data schema and data values. The various *minimal information* recommendations such as MIAME [9] and others[5] [10,11] are good examples. Ontology efforts such as the Gene Ontology [12] have made a major contribution to cross data linking. The existence of a shared ontology allows an integrator to combine multiple database records with at least some guarantee that the terms used by each resource correspond to each other. The National Center for Biomedical Ontologies[6] is a significant step in coordinating ontology development and annotation to better serve data pooling through shared controlled vocabularies [13]. There remains the challenge of ensuring and enabling widespread adoption. Most importantly, the tendency for political and theoretical wrangling [14] that so often accompanies ontology building must be minimised if practical progress is to be made and harm avoided.
- *The need for shared and stable access mechanisms.* The adoption of stable common formats, messages and protocols, and the publication of simple well defined APIs and query interfaces, greatly eases the plumbing together of data services. *Stability* is, however, the watch-word with interfaces. In 2007 BioMART [15] altered its interface four times, breaking any client software that used it. The National Center for Biotechnology Information[7] regularly alters its report format for its BLAST service with similar effects. A professional stance must be adopted by service providers to offer useful interfaces to their clients, and then to maintain them.
- *The need to adhere to standards.* The problems above are greatly alleviated by standardisation. Yet standards are boring. John Quackenbush describes them as "blue collar science". No-one will win a Nobel Prize for defining a workable format standard.
- *The need to explicitly state collection policies and governance.* It is crucial to have a clear understanding of the coverage and content of a collection, with minimum levels of quality for provenance and attribution. The governance policies of collections—responsibility for content, management of change, setting access criteria, security arrangements, licensing arrangements and so forth—are often confusing and ad hoc. Yet these are important considerations for any integrator.
- *The need to balance curation with exploitation.* The discipline's culture expects detailed human curation of data. Uniprot, for example, has scores of expert biologists working as curators, reading papers and making skilled judgments on how to describe a protein. This effort has often been poorly matched by the resources needed to exploit the data by making it easier to use.

In delivering solutions for these requirements, some fundamental social issues in bioinformatics must be borne in mind. Data integration is hard work. On the one hand, the scientific and political independence of the databases must be respected, and the freedom for rapid innovation celebrated. The creative enthusiasm of the bioinformatics community should not be dampened. On the other hand, the data held within them need to be unambiguously understood and easily integrated to address cross-database queries that span domain and organisational boundaries. There is

[3] http://www.uniprot.org.
[4] http://www.purl.org.
[5] http://www.nature.com/nbt/consult/index.html.
[6] http://bioontology.org.
[7] http://www.ncbi.nlm.nih.gov.

a need for balance, but there is fundamentally an overwhelming need to make data integration easier. Data integration is a prerequisite for much of today's biology, and as data production has been industrialised beyond a craft-based cottage industry, so must bioinformatics analysis.

4. Data integration regimes

The bioinformatics community is well aware of the need to support data integration so that scientists can "data surf". A wide variety of technologies, techniques and systems have been explored and exploited over the past 15 years: They vary on: the architectures they adopt, their reliance on manual or automated methods, and whether it is the data source or the integration system that bears the cost. They integrate on a range of common or corresponding touch-points: data values, names, identities, schema properties, ontology terms, keywords, loci, spatial-temporal points, etc. These touch-points are the means by which integration is possible, and a great deal of a bioinformatician's work is the mapping of one touch-point to another.

- They integrate by different mechanisms, for example: direct interlinking between database records, cross-database indexing, data exchange protocols, the merging of data sets against a common schema, the mapping of names and values between different data sets, aggregating all known data held on the same data instance, and interoperating different data-centric applications to build integration pipelines using workflows.
- They range from light touch solutions to more heavy-weight mechanisms (Fig. 1).

We now review some popular approaches to data integration and offer some observations.

Service oriented architectures, using technologies such as CORBA and Web Services [16], provide a uniform regime for "plumbing together" data resources that present themselves as services with programmatic interfaces. These technologies are not, however, integration mechanisms in themselves—once plumbed, the data have to be massaged and cleaned to make them fit together or conform to new schema. Progress has been mixed; even when efforts have been a technical success, they have not always been adopted. CORBA, for example, was felt by many to be too heavy-weight, leaving its early promise unfulfilled. Web Services have, however, had a greater penetration [17]. Fortunately, the importance of interacting with data through an interface other than a "point and click" Web page has now been widely recognised, and it is to be hoped that the days of simulating users and screen-scraping results are numbered. It is still the case, however, that Web Service interfaces are often poorly constructed, and poorly documented. On the one hand there are concerns that SOAP-based Web Services lack many features necessary for supporting integration and the technology is consequently becoming heavier.[8] On the other hand, REST (Representational State Transfer) is a simpler interaction model which, due to its ubiquity throughout the Web 2.0 movement, is gaining popularity [18]. The lesson is that of "Occam's Razor": any integration technology should be only as heavy as it needs to be and no heavier.

Link integration directly cross-references a data entry in a data source with another entry in another data source. Users follow the references. As these entries are usually presented as Web pages, the users surf across datasets by following hyperlinks. The approach leans heavily on ontology and identity authorities to enable the cross-referencing. Systems such as SRS [19], Entrez [20] and Integr8 [21] are portals and keyword indexing systems that maintain the interresource link network. This is still the most effective and widely used approach today, supported by the major service providers in the field. SRS still represents 40% of the EMBL-EBI traffic. This is, however, "Integration Lite". It is haphazard, and requires the cooperation of service providers to work well. It is vulnerable to name clashes, ambiguities and updates. It is really interlinking rather than integration, as the integration and interpretation is undertaken by some other mechanism—a person or another application. It is model independent, as there is no domain model to guide the integration.

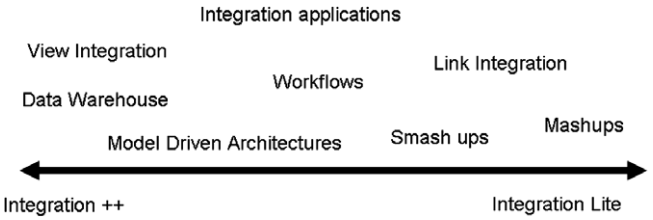

Fig. 1. A spectrum of data integration regimes.

Data warehousing, in contrast to link integration, is "Integration++". Data sets are extracted, cleaned and massaged into shape in order to be systematically combined into a (different) pre-determined domain or data model, usually devised by a third party, to be stored and queried as a single, free-standing, integrated resource. This is true data integration from many resources into one, long-standing resource. It is a popular approach—especially within sub-cultures for a particular species (e-Fungi [22], ATLAS [23], GIMS [24], Columba [25]) or for a discipline. It is also commonly found within enterprises for in-house data gathering. A range of toolkits—IBM's Websphere Information Integrator,[9] GMOD [2], BioMART[15,26], BioWarehouse [27]—flourish. However, this high gain approach has high pain. Warehouses require a pre-determined, encompassing model that should also remember the context of the warehoused data by tracking its source provenance. High initial activation costs mean that warehouses often represent an act of faith that they will be needed in the way they have been built. The model is often fixed, or hard to change, so that the user only gets what they are given and nothing more, despite changes in requirements or data. Once built, they have high maintenance costs to maintain synchrony with changes in their sources, especially as they are commonly decoupled from their data suppliers. They often struggle to adapt to source database churn and tinkering, which leads to brittle feeder wrappers. So, although popular, they find it hard to cope with a world in flux, and have consequently been likened to data mortuaries rather than warehouses. This is a symptom of data warehouses in general, and is not just confined to the Life Sciences: major industry commentators have observed that 30% of data migration projects fail[10] and 50% of data warehousing projects.[11]

View integration is another "Integration++" approach, but this time the data. is left in its source database and an environment is constructed that makes all the databases appear to be a single database. The result is a kind of "virtual warehouse" maintained by a mediator processor and a series of mappings from the integrating model to source (known as base) databases. Each database has a driver to extract data to match the mapping. The outcome is that the "content" of the view model is always fresh. This ambitious approach has been widely practised in the life sciences and is popular with database theorists and vendors. The database re-

[8] http://www.xs4all.nl/~irmen/comp/CORBA_vs_SOAP.html.

[9] http://www-03.ibm.com/industries/healthcare/doc/content/solution/939513305.html.

[10] http://www.standishgroup.com/.

[11] http://www.ncr.com/.

search community has developed theoretical approaches based on the mapping models—Global As View; Local As View and hybrids [28]. Examples of systems in the Life Sciences include BioZon [29], TAMBIS [30], Kleisli [31], Medicel Integrator,[12] IBM's Websphere Information Integrator, and ComparaGrid.[13] Both TAMBIS and ComparaGrid use an ontology as a global schema with query processing to transiently "fill" portions of that schema from distributed resources. Though intellectually appealing, this approach has not gained widespread adoption other than as an in-house enterprise solution. It carries the costs of a warehouse with respect to the development of a model; the models are hard to adapt; the drivers and mappings are often fat and brittle in the face of unreliable and dynamic resources; and the whole environment is complex, both for the developer and the user. They also tend to be as slow as the slowest source. Recent proposals on Dataspaces [32] argue that the way forward is best effort evidence-based integration and auto-generated mappings—echoing our view that just enough and just in time integration is desirable and practical.

Model-driven service oriented architecture is a version of view integration as practised by major projects such as caBIG [33], where a strongly typed data grid is decreed through the definition of Common Data Elements and a Common Vocabulary. All data resources and tools are obliged to adhere to this model in order to participate in the grid, exchanging data defined against the model. The benefits are that the system is designed as one rather than as many parts, which means it is generally only possible to achieve in tightly coupled systems with considerable penalties and/or incentives for the participating service providers. An attempt at a lighter touch approach has been tried by Gaggle [34]. Loosely coupled systems require an extensive battery of stable standards for data types and formats, and that shows little sign of happening in the bioinformatics domain; for example, EMBOSS reports over 20 different sequence formats recognised by its seqret[14] programme.

Integration applications such as Ensembl [22], Toolbus [35], Utopia [36] and Ondex [37] are built specifically to integrate data. They are not general integration systems as with views or warehouses; they are more specifically designed for a single application domain. Most have model-driven architectures. On the plus side these systems serve their single application domain well; on the downside they are specific and often difficult to extend. Some, such as the Data Playground [38], focus on enabling people to manually explore data mapping and relationships, observe their interactions and then automatically generate macros or workflows to replicate the integration pattern. These are hybrids of link integration mixed up with model-driven architectures, inheriting the weaknesses and strengths of both.

Workflows are a general technique for describing and enacting a series of linked processes. For data integration, the workflow coordinates a transient dataflow between data services and analytical tools. Workflows effectively systematically automate the in silico protocols that bioinformaticians have previously undertaken in ad hoc ways with PERL scripts. Rather than hiding the integration methods as in warehousing or views, all is exposed. Consequently, this technique could be called "Integration Self-help"—the workflow scripts are where the effort of integration takes place. Systems vary. Some, like InforSense[15] and Pipeline Pilot,[16] expect strong data type compliance so that the workflow effectively builds an on-the-fly data warehouse, or they presume a common data model. Others, like Taverna [39], have an open type system where the data passed between the workflow steps is "massaged" into shape by special "shim" processors and it is part of the responsibility of the workflow to build a data model if required. Workflow approaches have become extremely popular [11]. They are used to populate data warehouses and data views, yet are flexible and adaptable, and do not require the pre-existence of a single model. They presume that the data resources will be unreliable, so cater for service substitutions and faults. They typically do not hide the integration, so the evidence for data integration is exposed for scrutiny. Most maintain a provenance log [40] of their execution, giving an evidence trail of the integration. When used for scripting data chains, they are effectively automating the link integration approach. They are, however, not a universal panacea. Just as an experiment is hard to design, a workflow is hard to write, and the workflows are only as good as the services they link. Their greatest benefit is that a bioinformatician can make their own workflows (or re-use others) without a need to rely on one authority to do the integration. Workflows can be thought of a kind of mashup.

Mashups—of functional capability or content—are a Web 2.0 idea beginning to take hold in the sciences. Mashups provide a means to take data from more than one Web-based resource to make a new Web application. Data are taken from different Web Services with RESTful APIs or RSS-based syndication feeds, and "mashed" to provide new ways of combining and presenting information. An archetypal example takes a live content syndication feed on earthquake measurements and mashes it with Google Maps to present a new visualisation tool, overlaying sensor data on top of geospatial data.[17] By bringing information together in a new way we provision a new application with new functionality. The most common mashups—for example, news feed aggregators—are aggregation rather than integration; nonetheless they are incredibly useful. The attractiveness of a mashup is its Web delivery, its openness and its lightness; frameworks such as Microsoft's Popfly[18] and Yahoo! Pipes[19] make client-side development by the user straightforward. One of the first health care and life science mashup examples is the use of Google Earth to track the global spread of avian flu, reported in Nature News.[20]

Mashups emphasize the role of the user in creating a specific, light touch, on-demand integration, following the mantra of "just in time, just enough" design. Thus Web 2.0 mashups are built upon the existence of common *de facto* APIs alongside a collection of light-weight tools and techniques for rapid and agile deployment, so that small specific solutions can be built for particular problems. This can be contrasted to long, general engineering solutions that all too often do not meet a user's needs. Consequently, something light and quick is appealing, and it is easy to see how the idea of mashup would find traction within bioinformatics. Bioinformatics is a discipline where "just enough" and "just in time" are *de rigueur*—it is the biology that matters, not the engineering (for good or ill). Bioinformatics has always been a heavy user of the Web and now has many of its resources available as Web Services. The Distributed Annotation Service (DAS) [41] can rightly be thought of as an early form of a mashup service, and the Uniprot DASTY client combines 26 DAS Servers layering third party information about sequence annotations as tracks on the sequence [42].

Mashups are "Integration Lite++", where the prime activity is really *aggregation*. A mashup is transient; it lasts for as long as it is needed, and its interpretation is dictated by the mashup application rather than a model. Most other integration regimes aim for a degree of permanence, in that the integration persists in order that other applications can take advantage of it. Just as with link integration, mashups depend on having some kind of common

[12] http://www.medicel.com.
[13] http://www.comparagrid.org.
[14] http://emboss.sourceforge.net/apps/seqret.html.
[15] http://www.inforsense.com.
[16] http://www.scitegic.com/.
[17] http://earthquake.googlemashups.com/.
[18] http://www.popfly.com.
[19] http://pipes.yahoo.com.
[20] http://www.nature.com/news/2006/060105/full/news060105-1.html.

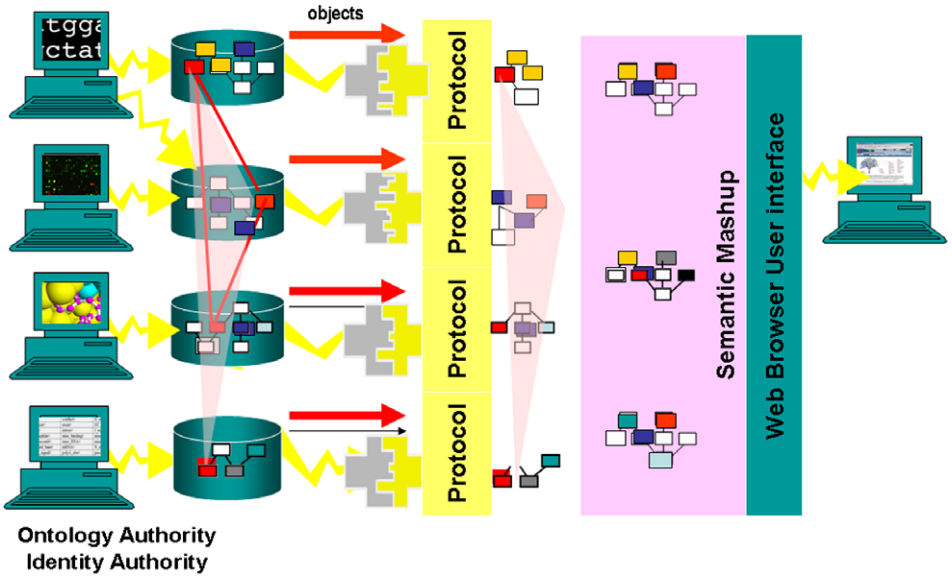

Fig. 2. A (semantic) mashup (based on figure by Cameron). Data are linked by common or discovered identities and shared annotations (tags) drawn from controlled vocabularies, managed by identity and ontology authorities. The data are accessed through simple programmatic APIs such as Atom, and aggregated through AJAX scripting in the browser based on these common identities and tags.

touch-point upon which to hang the various forms of information. It is possible to display sensor readings on a map only if there is some common means of locating where a sensor is on that map. Identity serves this role in bioinformatics, so mashups are just as vulnerable as everything else to identity clashes and concept ambiguities (Fig. 2).

5. The Semantic Web—mashups and smashups

This special issue of Biomedical Informatics focuses on the semantic extension of Web 2.0 ideas in biomedical data integration. First, we should examine the role of the Semantic Web in integration. We find the situation encouraging:

- *Publish data sets.* By publishing datasets as RDF (Resource Description Framework) on the Web [43] we overcome the structural boundaries of each resource's data model, effectively flattening the models. In contrast to the XML schema's tree structure, RDF has a graph-based structure. The latter provides a flexible, schema-less model that is responsive to change and inherently supports semantic descriptions of data. We can self-describe the exported data, potentially making it easier to be interpreted by an integration system, be that a workflow or a mashup.
- *Linked data sets.* By exposing, sharing and connecting pieces of data on the Semantic Web, it should be possible to navigate a "Web of Data" by following links from a data item within one data source to related data items within other sources, perhaps using a Semantic Web browser like PiggyBank [44]. The publishing of data as RDF, and the setting of RDF links between data items from different data sources, is the foundation. RDF links could also be crawled and indexed by Semantic Web search engines like Sindice.com [45], which could provide sophisticated search and query capabilities over the crawled data. As query results are structured data and not just links to HTML pages, they can be used within other applications such as DBpedia [46]. The W3C Semantic Web Education and Outreach (SWEO) Interest Group Linking Open Data community project[21] is currently developing best practice proposals for Web access to Linked Data, emphasising the use of conventional Web technologies such as URIs and HTTP, alongside mandates such that resource URIs should be de-referenced and yield metadata about the resource.[22]
- *Ontologies.* By using RDFSchema (RDFS) and OWL (Web Ontology Language) we have the means to create ontologies covering a spectrum of richness, from simple taxonomies as in Simple Knowledge Organisation Systems (SKOS)[23] to full-blown knowledge models.
- *Mapping models.* By using the linking and reasoning capabilities of the Semantic Web we should be able to build mappings between ontologies and data entities.
- *Semantic enrichment.* By annotating resources, using stand-off annotation (where content and the annotations associated with content are separated) we are able to describe services and workflows to aid in service discovery [47] or workflow assembly [38]. Embedded microformats and XML-based semantic overlays such as RDFa[24] enable us to enrich Web content with semantics. We should also be able to enrich data with its provenance, its context or the evidence supporting its correspondence with other data.
- *Supportive metadata.* By using RDF we can represent the provenance trails of our integration systems in a flexible and extensible manner.

So how is the Semantic Web aiding data integration? There are plenty of examples of ontology development using OWL, although the rivalry with the Open Biomedical Ontologies (OBO) format is a distraction that we hope will soon be set aside. RDF has been proposed as a flexible model for standard data collection [48], and its capability of graph merging is a potential boon to integration, since once a URL is found in common between two RDF graphs, those graphs can be merged into one aggregated data structure. Workflow systems such as Taverna have made use of RDF and OWL to annotate their systems and record provenance logs for the integration outcomes [38]. BioMOBY [49,50] is an example of a system

[21] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[22] http://www.w3.org/DesignIssues/LinkedData.html.

[23] Simple Knowledge Organisation Systems (SKOS) http://www.w3.org/2004/02/skos/.

[24] http://www.w3.org/TR/xhtml-rdfa-primer/.

where service annotations are used to produce an RDF graph that maps the links between datasets. The Sealife project [51] uses semantic annotation on Web pages to provide an enriched link experience when browsing Life Science Web pages in a browser. We have examples of building RDF warehouses such as YeastHub [52], and RDF on-demand caches such as BioDASH [53]. However, these have the same issues as all warehouses, and we have little practical experience with how the Semantic Web can help with source churn, though in principle it ought to be able to cope more effectively with schema changes. Although interlinked datasets with common vocabularies are not yet widespread in the general Web community [54] they are prevalent in the Life Sciences, and form the basis of the W3C Healthcare and Life Sciences Interest Group's (HCLSIG) Semantic Web demonstrators [13,55]. These demonstrators also exploit the graph merging capabilities of RDF.

The ultimate goal of a Semantic Web for Life Sciences is not to create many separate, non-interacting data warehouses, but rather to create a single Web of biological data and knowledge that can be crawled and queried, similar to the existing Web. To achieve this vision, we must go beyond building pre-compiled RDF warehouses [56]. For this we turn to a combination of the Linked Data vision we sketched above and the exciting possibilities of semantic mashups, or "smashups" (Fig. 2). The Nature News avian flu mashup, referred to previously, which linked up UN Food and Agricultural Organization location data with Google Earth, ran into problems with mismatches between the two sources as they differed on the conceptualisation underlying their location coordinate systems. A mashup is only as good as its shared data points upon which the mashup is made; clearly shared Semantic Web ontologies will help by enabling mashups to share data and interoperate by using shared or mapped semantic terms as mash-points, and by making much more explicit the potential for conceptual ambiguity and misunderstandings. Moreover, expressively defined knowledge on the Web will enable mashups to better discover and access existing information. Semantic knowledge might also encourage the innovation of non-map-based mashups.

For smashups to work we must:

- persuade service providers to offer simple and stable RESTful APIs that can be used by third parties in Web browsers;
- persuade service providers to: publish their data as RDF; build systems that export legacy data to the Web as RDF; and expose SPARQL[25] endpoints to datasets so that they can be queried using Semantic Web search engines
- encourage the community to tag content with shared semantic terms that can be used as "smashup touch-points" or used to clarify the intended semantics of data
- persuade the community to seriously tackle the profound problem of entity identity.

Much of the success of mashups lies in the underlying attitude of the approach: its simplicity; its user participation model; and its perpetual beta production ethos [57]. Straightforward tagging models have the virtue of simplicity; the introduction of richer semantic models runs the risk of added complexity which could compromise the mashup philosophy, or make it only accessible to specialist developers of integration systems. Toolkits such as Popfly and Yahoo! Pipes have made mashup development accessible to a wider community; similar tools that cope with RDF and OWL would be a real asset.

[25] SPARQL is the W3C proposed query language for RDF http://www.w3.org/TR/rdf-sparql-query/.

6. Conclusion

If the bioinformatics community could become better organised on only one topic, then it should be addressing the issue of identity and naming. This would have the most profound consequence for data integration. Projects such as Bio2RDF [58] are a step towards the provision of real time translation and harmonization of identifiers over bioscience datasets, but have yet to gain real traction. The failure to address identity will be the most likely obstacle that will stop mashups, or any other technology or strategy, becoming an effective integration mechanism. Many of the services in workflows are identity transformation mappings. Much of the work in view and data warehouse integration is finding out if two different data entries, and the entities they describe, are the same thing. This is not a call for some coherence in the functional naming of genes and proteins; it is simply pleading that some semantic-free unique identifier be made for the basic entities being described in bioinformatics data resources.

The W3C HCLSIG has a real and important role to play: to grasp the nettle of identity management and to show how, using lightweight semantic techniques, we can rapidly aggregate data just in time and just when it needs to be, by the user and for the user.

Acknowledgments

We gratefully acknowledge: Graham Cameron for the inspiration for Fig. 2; and Rodrigo Lopez, Rolf Apweiler, Mark Wilkinson, Matt Pocock, Duncan Hull, Suzanne Embury, Peter Li, Kei Cheung, Susie Stephens and David De Roure for many suggestions, comments and insights. We also thank the anonymous reviewers for their helpful comments.

References

[1] Galperin MY. The molecular biology database collection: 2007 update. Nucleic Acids Res 2007;35(1):D3–4.
[2] Stein L. Creating a bioinformatics nation. Nature 2002;417(6885):119–20.
[3] Baxevanis AD. The molecular biology database collection: an updated compilation of biological database resources. Nucleic Acids Res 2001;29(1):1–10.
[4] Merali Z, Giles J. Databases in peril. Nature 2005;435(7045):1010–1.
[5] Stein LD. Integrating biological databases. Nat Rev Genet 2003;4(5):337–45.
[6] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The universal protein resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34:D187–91.
[7] Kanehisa M, Goto S, Kavashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32:D277–80.
[8] Clark T, Martin S, Liefeld T. Globally distributed object identification for biological knowledge bases. Brief Bioinform 2004;5(1):59–70.
[9] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)[mdash]toward standards for microarray data. Nat Genet 2001;29:365–71.
[10] Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotech 2007;25(8):894–8.
[11] Taylor CF, Paton NW, Lilley KS, Binz P-A, Julian RK, Jones AR, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotech 2007;25(8):887–93.
[12] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[13] Stephens S, LaVigna D, DiLascio M, Luciano J. Aggregation of bioinformatics data using semantic web technology. Web Semant 2006;4(3):216–21.
[14] Goble C, Wroe C. The montagues and the capulets. Comp Funct Genomics 2004;5(8):623–32.
[15] Kasprzyk A, Keefe D, Smedley D, Darin L, William S, Craig M, et al. EnsMart: a generic system for fast and flexible access to biological data. Genome Res 2004;14:160–9.
[16] Gisolfi D. Web Services Architect Part 3: Is Web services the reincarnation of CORBA? IBM Developer Works 2001 [cited October 2007]. Available from: http://www.ibm.com/developerworks/webservices/library/ws-arc3/.
[17] Neerincx PBT, Leunissen JAM. Evolution of web services in bioinformatics. Brief Bioinform 2005;6(2):178–88.
[18] Prescod P. REST and the Real World. XMLcom 2002 [cited October 2007]. Available from: http://webservices.xml.com/pub/a/ws/2002/02/20/rest.html.

[19] Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. Methods Enzymol 1996;266:114–28.
[20] Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. Methods Enzymol 1996;266:141–62.
[21] Kersey P, Bower L, Morris L, Horne A, Petryszak R, Kanz C, et al. Integr8 and genome reviews: integrated views of complete genomes and proteomes. Nucl Acids Res 2005;33(Suppl. 1):D297–302.
[22] Hedeler C, Wong HM, Cornell MJ, Alam I, Soanes DM, Rattray M, et al. e-Fungi: a data resource for comparative analysis of fungal genomes. BMC Genomics 2007;8:426.
[23] Shah S, Huang Y, Xu T, Yuen M, Ling J, Ouellette BFF. Atlas—a data warehouse for integrative bioinformatics. BMC Bioinformatics 2005;6(1):34.
[24] Cornell M, Paton NW, Wu S, Goble CA, Miller CJ, Kirby P, et al. GIMS—a data warehouse for storage and analysis of genome sequence and functional data. In: Proceedings of the 2nd IEEE international symposium on bioinformatics and bioengineering. Bethesda, MD, USA: IEEE Computer Society; 2001.
[25] TriSzl S, Rother K, Muller H, Steinke T, Koch I, Preissner R, et al. Columba: an integrated database of proteins, structures, and annotations. BMC Bioinformatics 2005.
[26] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005;21(16):3439–40.
[27] Lee T, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert D, Tenenbaum J, et al. BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics 2006;7(1):170.
[28] Alon YH. Answering queries using views: a survey. VLDB J 2001;10(4):270–94.
[29] Birkland A, Yona G. BIOZON: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinformatics 2006;7(1):70.
[30] Baker P, Goble C, Bechhofer S, Paton N, Stevens R, Brass A. An ontology for bioinformatics applications. Bioinformatics 1999;15(6):510–20.
[31] Davidson SB, Crabtree J, Brunk BP, Schug J, Tannen V, Overton GC, et al. K2/Kleisli and GUS: experiments in integrated access to genomic data sources. IBM Syst J 2001;40(2):512–30.
[32] Alon H, Michael F, David M. Principles of dataspace systems. In: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Chicago, IL, USA: ACM; 2006.
[33] Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, et al. caCORE: a common infrastructure for cancer informatics. Bioinformatics 2003;19(18):2404–12.
[34] Shannon P, Reiss D, Bonneau R, Baliga N. The Gaggle: an open-source software system for integrating bioinformatics software and data sources. BMC Bioinformatics 2006;7(1):176.
[35] Eckart JD, Sobral BW. A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework. Omics 2003;7(1):79–88.
[36] Pettifer SR, Sinnott JR, Attwood TK. UTOPIA: user friendly tools for operating informatics applications. Comp Funct Genomics 2004;5(1):56–60.
[37] Koehler J, Rawlings C, Verrier P, Mitchell R, Skusa A, Ruegg A, et al. Linking experimental results, biological networks and sequence analysis methods using ontologies and generalised data structures. In Silico Biol 2004;5(5).
[38] Gibson A, Gamble M, Wolstencroft K, Oinn T, Goble C. The data playground: an intuitive workflow specification environment. In: e-Science 2007—third IEEE international conference on e-science and grid computing. Bangalore, India: IEEE Computer Society; 2007. pp. 59–68.
[39] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004;20(17):3045–54.
[40] Goble C, Wolstencroft K, Goderis A, Hull D, Zhao J, Alper P, et al. Knowledge discovery for biology with taverna: producing and consuming semantics in the web of science. In: Baker CJO, Cheung K-H, editors. Semantic web: revolutionizing knowledge discovery in the life sciences. USA: Springer; 2007. p. 355–95.
[41] Dowell R, Jokerst R, Day A, Eddy S, Stein L. The distributed annotation system. BMC Bioinformatics 2001;2(1):7.
[42] Jones P, Vinod N, Down T, Hackmann A, Kahari A, Kretschmann E, et al. Dasty and UniProt DAS: a perfect pair for protein feature visualization. Bioinformatics 2005; May 19, 2005:bti506.
[43] Antoniou G, van Harmelen F. A semantic web primer. MIT Press; 2004.
[44] Huynh, D., Mazzocchi, S., Karger, D. Piggy bank: experience the semantic web inside your web browser. In: Gil Y, Motta E, Benjamins, VR, Musen, MA, editors. 4th international semantic web conference (ISWC 2005), 2005. Galway, Ireland: Springer Berlin/Heidelberg; 2005. pp. 413–30.
[45] Tummarello G, Delbru R, Oren E. Sindice.com: weaving the open linked data. In: Aberer K, Choi K-S, Noy NF, Allemang D, Lee K-I, Nixon LJB, et al., editors. 6th international semantic web conference and 2nd asian semantic web conference (ISWC/ASWC2007), 2007. Busan, South Korea: Springer Berlin/Heidelberg; 2007. p. 547–60.
[46] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. DBpedia: a nucleus for a web of open data. In: Aberer K, Choi K-S, Noy NF, Allemang D, Lee K-I, Nixon LJB, et al., editors. 6th international semantic web conference and 2nd asian semantic web conference (ISWC/ASWC2007), 2007. Busan, South Korea: Springer Berlin/Heidelberg; 2007. p. 715–28.
[47] Lord P, Bechhofer S, Wilkinson M, Schiltz G, Gessler D, Hull D, et al. Applying semantic web services to bioinformatics: experiences gained, lessons learnt. In: McIlraith SA, Plexousakis D, van Harmelen F, editors. Third international semantic web conference (ISWC 2004), 2004. Hiroshima, Japan: Springer; 2004. p. 350–64.
[48] Wang X, Gorlitsky R, Almeida JS. From XML to RDF: how semantic web technologies will change the design of /'omic/' standards. Nat Biotechnol 2005;23:1099–103.
[49] Wilkinson M, Schoof H, Ernst R, Haase D. BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet exemplar case. Plant Physiol 2005;138(1):5–17.
[50] Wilkinson MD, Links M. BioMOBY: an open source biological web services proposal. Brief Bioinform 2002;3(4):331–41.
[51] Bechhofer S, Yesilada Y, Stevens R, Jupp S, Horan B. Using ontologies and vocabularies for dynamic linking. IEEE Internet Comput, in press.
[52] Cheung K-H, Yip KY, Smith A, deKnikker R, Masiar A, Gerstein M. YeastHub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics 2005;21(Suppl. 1):i85–96.
[53] Neumann EK, Quan D. Biodash: a semantic web dashboard for drug development. Pac Symp Biocomput 2006:176–87.
[54] Ding L, Finin T. Characterizing the semantic web on the web. In: Cruz IF, Decker S, Allemang D, Preist C, Schwabe D, Mika P, et al., editors. 5th international semantic web conference (ISWC 2006), 2006. Athens, GA, USA: Springer; 2006. p. 242–57.
[55] Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the semantic web. BMC Bioinformatics 2007;8(Suppl. 3):S2.
[56] Good BM, Wilkinson MD. The life sciences semantic web is full of creeps! Brief Bioinform 2006;7:275–86.
[57] Musser J, O'Reilly T. Web 2.0 Principles and Best Practices: O'Reilly Media; 2006.
[58] Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge system health care and life sciences data integration for the semantic web. Banff, Canada. Available from: http://bio2rdf.org/2007.