BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 6. 562–575
Advance Access published on 3 October 2011
doi:10.1093/bib/bbr051

Knowledge sharing and collaboration in translational research, and the DC-THERA Directory

Andrea Splendiani, Michaela Gündel, Jonathan M. Austyn, Duccio Cavalieri, Ciro Scognamiglio and Marco Brandizi

Submitted: 13th May 2011; Received (in revised form): 2nd August 2011

Abstract

Biomedical research relies increasingly on large collections of data sets and knowledge whose generation, representation and analysis often require large collaborative and interdisciplinary efforts. This dimension of 'big data' research calls for the development of computational tools to manage such a vast amount of data, as well as tools that can improve communication and access to information from collaborating researchers and from the wider community. Whenever research projects have a defined temporal scope, an additional issue of data management arises, namely how the knowledge generated within the project can be made available beyond its boundaries and life-time. DC-THERA is a European 'Network of Excellence' (NoE) that spawned a very large collaborative and interdisciplinary research community, focusing on the development of novel immunotherapies derived from fundamental research in dendritic cell immunobiology. In this article we introduce the DC-THERA Directory, which is an information system designed to support knowledge management for this research community and beyond. We present how the use of metadata and Semantic Web technologies can effectively help to organize the knowledge generated by modern collaborative research, how these technologies can enable effective data management solutions during and beyond the project lifecycle, and how resources such as the DC-THERA Directory fit into the larger context of e-science.

Keywords: *semantic web; ontology; immunology; eScience; data integration*

INTRODUCTION

Biomedical research is increasingly reliant on large collections of data and knowledge that require computational approaches for their management and analysis [1]. Deriving knowledge from large amounts of data requires it to be properly organized so that relationships among data elements are understood and put into the context of current knowledge [2]. This is a particularly challenging task in the biomedical domain where information is complex and often relates data with multiple levels of granularities and that pertain to different disciplines [3]. In recent years we have witnessed the development of tools and techniques, the focus of which has evolved from the basic storage and retrieval of data to more versatile tools that enable the integration of

Corresponding authors. Dr Andrea Splendiani, IntelliLeaf Ltd, Cambridge CB1 3UF, UK. Tel.: +44 (0) 1223 853544; Fax +39 02 42108159; E-mail: andrea@leafbioscience.com and Duccio Cavalieri, Dipartimento di Farmacologia, Università di Firenze, Viale Pieraccini 6, 50139, Firenze, Italia Tel: +39.055.4271327; Fax: +39.055.4271280; Email: duccio.cavalieri@unifi.it

Andrea Splendiani is a Founder of IntelliLeaf ltd and a Senior Bioinformatic Scientist at Rothamsted Research. His research interests are in Semantic Web technologies, biomedical ontologies and Systems Biology.

Michaela Gündel holds an MSc degree in Life Science Informatics. Her main research interest lies in semantic knowledge management, ontology development and their application to the life sciences.

Jonathan M. Austyn is Professor of Immunobiology at the University of Oxford and Project Coordinator for the DC-THERA Network. His research interests are the immunobiology and therapeutic applications of dendritic cells.

Duccio Cavalieri is Professor at the Faculty of Pharmacy, University of Florence, applies computational biology to the investigation of the signaling networks of the dendritic cells through pathway analysis of functional genomics data sets.

Ciro Scognamiglio is a Freelance Web Professional based in Paris. His main expertise is in Software Development and Unix System Administration.

Marco Brandizi is a Founder of IntelliLeaf Ltd and a software engineer at the European Bioinformatics Institute, where he works on the management of microarray data and other multi-omics data.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

heterogeneous data and their annotation through standard terminologies. More recently, we have also seen the emergence of tools that support the social aspect of collaborative and interdisciplinary research.

As biomedical research started to become an information-intensive discipline, the focus of bioinformatics research initially was the creation of data-specific databases to store and enable searches over a growing quantity of data such as sequences [4] and gene expression [5, 6], or more complex information such as pathways [7] and scientific knowledge represented in the literature [8].

However it soon became clear that a proper metadata framework to annotate data was essential for making sense of the information stored in these databases [9, 10]. For instance, the functional genomics community pioneered the development of shared and computable terminologies (ontologies) to define experimental conditions [5], which resulted in the construction of the Ontology for Biomedical Investigation (OBI) [11].

Biomedical research often requires the integration and analysis of different types of information in a biological system, which is a complex task, as this information is often stored in different databases and represented differently. As a consequence, much research has been carried out on how best to manage, interrelate and interrogate biomedical data [12–14]. The task of 'Data integration' poses both technical and semantic challenges, which are often interconnected. The technology for relating information artifacts has evolved from the linking of flat files, through specialized software solutions [15], to web-based information systems that capitalize on the use of ontologies to provide distributed knowledge bases [16]. Underpinning these technologies are tools that allow the composition of data and services, which in turn have evolved from middleware such as CORBA [17] to web services, orchestrated web services [18] and advanced user interfaces and interactive environments [19]. We are now witnessing the convergence of solutions that merge ontology-enabled web services with the declarative nature of the web [20].

Beyond the techniques that have evolved to relate and exchange information across distinct databases, there is a need for the definition of common 'languages' to describe integrated information. When data integration was carried out within homogenous research communities, those languages could rely on a shared understanding of their concepts and the semantic challenges of data integration were addressed through the definition of 'exchange languages' [21, 22].

As research became increasingly interdisciplinary, however, the necessity for a common understanding of terms across different disciplines prompted the development of ontologies, such as the Gene Ontology (GO) [23], the success of which has led to the development of coherent ontology libraries, such as the Open Biomedical Ontologies (OBO) collection [24].

The definition of these biomedical ontologies has evolved both in its ontological foundations, with the commitment to common upper ontologies such as the Basic Formal Ontology (BFO) [25], and in its representation, which has become increasingly logic based, via the adoption of ontology definition languages such as the Web Ontology Language (OWL) [26–29]. Ontologies now comprise the backbone of biomedical informatics, with dedicated institutions such as the National Center for Biomedical Ontology (NCBO) [30] and resources such as the BioPortal [31]. Use of ontologies is not limited to the annotation of databases [32]: ontologies have provided a significant contribution to high-throughput data analysis and increasingly are seen as a device to make scientific literature more machine processable. Hence, they enable researchers to make better use of the increasing amount of knowledge available in this format. The gap between databases and scientific literature is narrowing [33–35].

From a wider perspective, the definition of ontologies and the increasing relevance of web-based technologies are part of a larger evolution of science (and knowledge creation in general), characterized by a computationally enabled social dimension [36]. This evolution has far-reaching consequences that touch the role of the public in scientific research, for instance, through 'crowd sourcing' [37] and through ownership of information [38].

So far, development of web-based resources that represent information through shared computable languages has focused on 'primary products' of research, such as datasets and literature. There are reasons for developing similar resources that focus on the research process itself. Research is often organized into projects that involve a network of collaborating participants who need to communicate and share intermediate results, best practices and, in general, their know-how. This necessity for

communication and knowledge sharing is not dissimilar to the needs of the biomedical community at large, and sometimes such networks have adopted knowledge management solution that mimic the functionalities of public repositories [39]. More often, they have relied on tools commonly used for project information and communication, such as wikis, mailing lists or content management systems (CMS).

It is useful to devise (web-based) resources that bring together these types of tools and that can both support the project-related activities of research communities and at the same time integrate their information with that of distributed repositories. Too often, metadata and data curation are left as a final step of research, causing precious information, useful for qualifying the output of research, to be lost [40]. Furthermore, there is a clear potential for error detection and the reduction of duplicate efforts.

Such resources can facilitate the sharing of much more relevant and useful information than can traditional methods such as publications: small facts and negative results can be published via web-based systems more widely and efficiently than via scientific literature, and they can be managed by means of computational systems that can provide credit for their generation [41].

Finally, such resources can integrate the social side of scientific research with the information that it generates, thus improving communication and collaboration among researchers. This is particularly true for the 'long tail' of researchers who share some specific interest, but who otherwise might be remote from the core community.

DC-THERA AND THE DC-THERA DIRECTORY

DC-THERA [42] is a European Network of Excellence (NoE) established under the European Commission's Sixth Framework Program, which has integrated many researchers and clinicians, working collectively on basic scientific and therapeutic aspects of dendritic cells (DC), a topic central to immunology. The network has brought together at least 32 partners and 38 associated partners, from 18 different European countries. It is a typical example of a translational and distributed research project, which has prompted the need for a computational, community-based approach to manage a wide range of heterogeneous information. The organization of information in DC-THERA poses additional challenges, since research focusing primarily on DC requires a characterization of resources by their cell-type specificity that often transcends the characterization provided by generic tools and information resources. The nature of DC-THERA as a research project also highlights issues about the way the information generated can be maintained after the project has ended, and how such information can be absorbed or re-used by other efforts that can emerge from the DC-THERA and from the wider community.

In this article we describe the DC-THERA Directory ([43, 44], hereafter called also 'the Directory'). This is a web-based knowledge management system, initially designed to address the collaborative and sharing needs of the DC-THERA community. The Directory focuses on the 'network knowledge', which is the set of technical resources, research expertise, personnel and their relationships that make up the core of a NoE and similarly organized communities.

The design of the DC-THERA Directory addressed three main goals.

First, to provide an information gateway for the DC research community that enriches proprietary and public information through annotations and search functions and provides focused information set for consumption by its researchers and other computational systems.

Second, to represent information in the Directory through languages and terminologies that are 'compatible' with the biomedical information ecosystem.

And, last, to maximize the 'integrability' of the represented information with external resources, so as to maximize its usefulness and visibility, beyond the boundaries of the specific research network that was initially served.

THE DC-THERA DIRECTORY AS AN EXAMPLE OF AN E-SCIENCE PORTAL

The DC-THERA Directory is a public web site [45] that provides information on research assets available within the DC-THERA community and, at the same time, integrates external resources to provide a coherent access point for researchers. Like other e-science resources, the Directory relies on annotations through ontologies and standard languages to provide advanced search and organization functions.

Table 1: The list of ontologies that are included in the DC-THERA Ontology

Ontology	Domain	Usage in DC-THERA Directory
Ontology for biomedical investigations (OBI) [11]	Meta-information for Biomedical experiments	Biomaterials, protocols, data sets, documents, tools and methods
Dendritic cell ontology [87]	Cell-type annotation	Biomaterials, data sets, protocols
Cell-type ontology (CL) [88]	Cell-type annotation	Biomaterials, data sets, protocols
Experimental factor ontology (EFO) [89]	Meta-information for microarray and -omics experiments	data sets
Microarray experimental conditions (MGED) [90]	Meta-information for microarray experiments	data sets
Chemical entities of biological interest (CHEBI) [91]	Annotation of bio-molecules and administered compounds/drugs	Biomaterials, data sets
Foundational model of anatomy (FMA) [88]	Annotation of biomaterials	Biomaterials, data sets, protocols
NCBI taxonomy [92]	Classification of organisms	Biomaterials, data sets, protocols

It also offers functions for editing information and for managing its privacy. At the time of writing, the Directory provides summary information on 237 data sets, 79 protocols, 524 biological materials, 122 laboratory tools (which include both equipment and consumables), 79 organizations and 328 persons. In addition, it integrates internal and external microarray repositories and provides literature and pathway analysis services.

Annotation, ontologies and standards

The DC-THERA Directory is organized such that one research asset corresponds to one information item in the Directory. Each information item is annotated via a brief textual description, a type, a set of attributes and its relationships to other resources. Complex resources are represented in more detail, as is the case for protocols, where protocol steps, their order and their requirements and results are represented explicitly. It is worth mentioning that a relevant feature of the Directory is its annotation of resources, protocols, data sets and eventually tools in terms of the specific cell type to which they relate, namely DC. This is of particular importance in immunology, as the interplay of different cell types is key aspect of the immune system. As shown in the next section, the Directory addresses the specificity of DC biology by annotating cell-type specific reactions and reagents and by using more general types to cross-connect the results stemming from the interaction of different cell types.

Most of the types, attributes and relations used in the Directory are drawn from ontologies of the OBO family, which makes the knowledge representation of the Directory contents interoperable with other related biomedical knowledge. To obtain a seamless and simplified framework for using these existing ontologies to annotate the Directory resources, a DC-THERA application ontology has been defined, using the standard ontology language, OWL. This ontology mostly extends the OBI, and integrates several other OBO ontologies (Table 1) in a way that suits the Directory annotation purposes. Moreover, several relationships and classes have been defined in order to achieve a balance between ontological precision and usability. For instance, a relationship is provided to link a cell culture to an ontology concept that represents the type of cells comprising that culture. While this is presented as a 'cell culture X of type Y', the relationship is actually a short-cut for the more correct statement: 'the cell culture X is a population of cells such that each cell has the property of being an instance of the type Y'. Not only does this short-cut simplify the editing tasks for the curator end-user, it also leaves room to derive the correct inference in an ontological framework (e.g. by means of rules). Another more trivial example is the use of an ad-hoc relation 'is-knowledgeable-about', defined to cover a range of relationships that could not be specified further (e.g. 'has produced the bio-material in the laboratory' or 'is an expert in the protocol'). A brief overview of the top-level classes and relationships defined in the DC-THERA ontology is provided in Figure 1.

User interaction

The Directory fulfills its role of information gateway for researchers by providing interactive search, result inspection and editing functions. In the Directory these functions often capitalize on the annotation of its information via ontologies and shared relations.

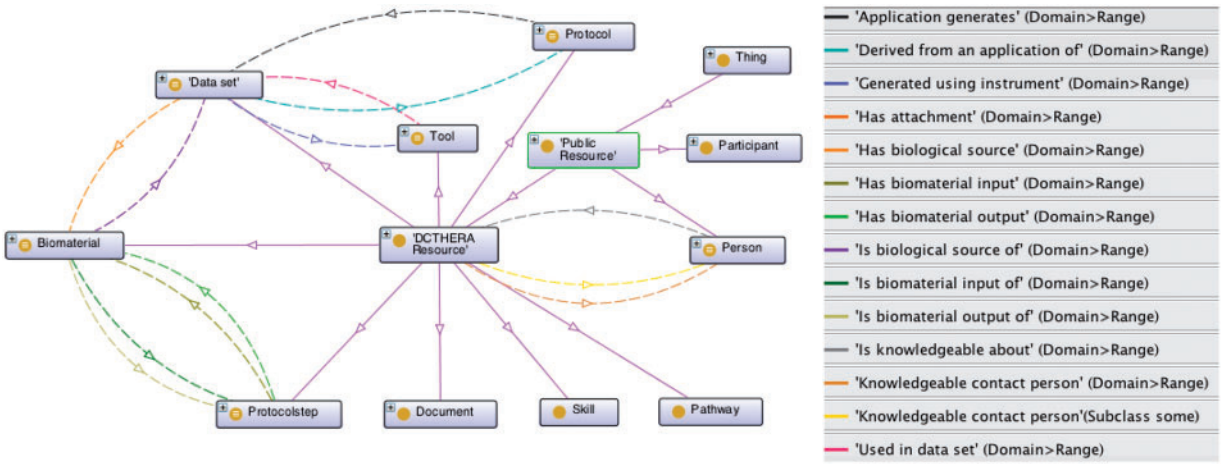

Figure 1: Extract from the DC-THERA Ontology. Some of the top-classes and relationships that are part of the DC-THERA Ontology are represented. The diagram makes use of labels in place of identifiers for readability. 'DC-THERA Resource' encompasses research assets available within DC-THERA, while 'Public Resource' is a more generic class that includes Participants and Persons (both of these classes are at a level of abstraction that is above what is presented to the user).

Search

The Directory provides different ways of accessing information. Ontologies are used as taxonomical indexes to organize and access its content, a significant case being access by cell type (or bio-material). In addition, the Directory also provides a simple 'Google-like' query interface that assists the user dynamically by providing predictive suggestions while typing. Keywords entered in the search forms are expanded in their synonyms and morphological variants and are then used to match types and text in the Directory, as well as to query external services. Results of a query are presented as a list with a brief description, where a color code identifies text matching exact terms or synonyms.

In some cases, the Directory tracks the user behavior to restrict free text queries so as to provide more 'intuitive' results. As an example, if the user is performing a text query while observing the list of data sets retrieved through an access via data type (taxonomy), the query is limited to data sets in the Directory, external resources that contain information on data sets (e.g. ArrayExpress) and other entries in the Directory that are related to the results found.

The Directory relies on ontologies to expand the scope of queries from more generic to more specific terms. For instance, a query for 'Leukocyte' will query the Directory also for all known sub-types of 'Leukocyte', as defined in the DC-THERA Ontology and as presented in the Directory in the bio-material taxonomy. Queries can then be used to extract the content of the Directory (e.g. reagents, data sets, protocols), and hence navigate its content, with the desired level of generalization in the specification of cell types.

Contextualization

For each resource, the Directory presents a 'resource-centric' view that provides a description of the resource and its context in the Directory: a brief overview of which other entries in the Directory relate to the resource in question, and how (Figure 2).

The description of a resource is generally in the form of a short piece of text and a list of features, organized in property/value pairs. Depending on the resource type, additional detail can be presented. This is the case for protocols, where the detailed description of their workflow is provided via a graph.

From the 'resource-centric' view, a user can easily identify other relevant resources in the same context and navigate the content of the Directory to which they relate. For example, a user can navigate from the description of a data set to the analysis protocol that was used to generate this data-set (where 'generate' is the property linking the two), then to a specific tool 'used in' the protocol, and from there to a member of the DC-THERA Network 'knowledgeable about' that tool.

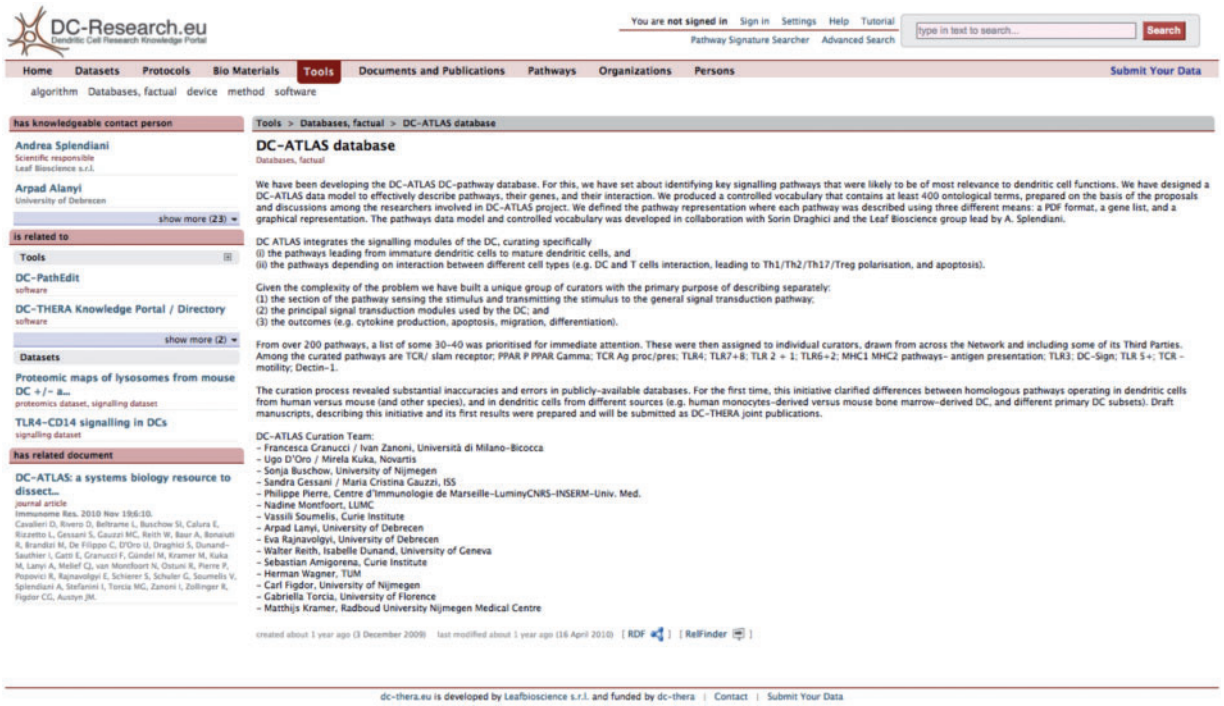

Figure 2: An example of a 'resource-centric' view in the DC-THERA Directory. The information page shown corresponds to the resource 'DC-ATLAS' (URL: http://dc-research.eu/tool/101). The classification of this resource as a tool is shown in the upper part of the page. Relationships with other resources in the Directory (context) are shown in the left bar. Note that the category 'Tools' is highlighted: this is an indication of the current context, and searches via the search box are performed on tools and related resources.

Privacy and annotation

Users can annotate data in a simple way. If they have sufficient permissions, they can annotate entities displayed in the 'resource-centric' view with properties/values, or with relationships to other objects. The Directory guides the user by proposing relationships or attributes that are desirable or sensible for the type of entity described. Users can then decide whether to make the created information public or whether to restrict it to a set of participants. The Directory supports a data access model where users can belong to different groups with different roles. For each group, and depending on their role, users may be able to read, edit, delete or even supervise curation of the entries assigned to it. A similar access model is described in [46].

DATA MANAGEMENT, INTEGRABILITY AND THE INFORMATION LIFECYCLE

In the design of the Directory we have addressed data management issues that relate to the lifecycle of a research project.

One goal has been to guarantee the longevity of the information in the Directory beyond the duration and the scope of the project. This has posed two problems: an economic problem, since resources to maintain the system cannot be guaranteed beyond the duration of the project, and a 'usability' problem, to ensure that the information in the Directory can be found easily, transported and manipulated with other systems to maximize 're-use' of the information generated within the network.

To address these issues, we have leveraged on the formal annotation of the entries in the Directory by making them available through standard technologies of the Semantic Web framework [47], such as RDFa [48] or SPARQL [49], details on which will be briefly presented later.

We present a few examples here that show how adoption of these technologies can improve the data management lifecycle.

Reachability

Most of the information in the Directory has been made publicly available on the web, after an initial phase in which access was restricted to DC-THERA

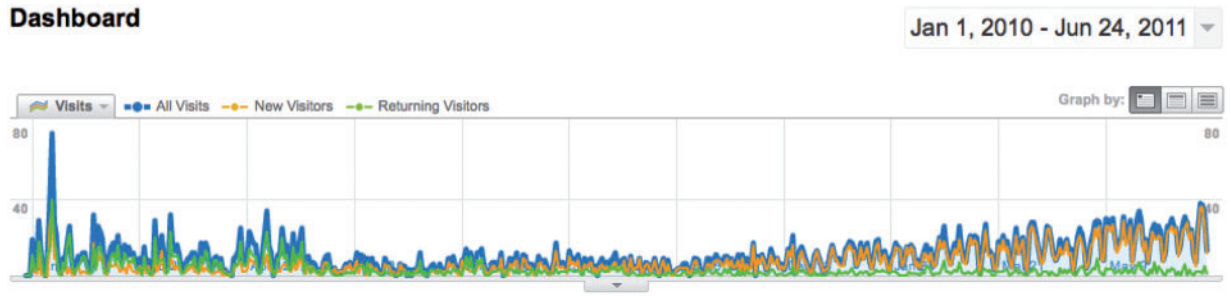

Figure 3: Web access data for DC-RESEARCH.EU from 1st December 2010 to 14th January 2011. Total figures are reported in blue, returning users in orange and new users in green. The figures show a drop in access towards the end of the DC-THERA project, and a slow and steady reprise afterwards, arguably corresponding to a shift in usage from a project-oriented tool to a generic web resource. Reported values exclude computational access via a SPARQL end-point and access through a replicated platform (cf. 'Portability and long-term persistence' section).

participants. The opening of the Directory to the public was motivated by its potential value to the wider scientific community, including the potential for spawning new collaborations and ideas, due to the links that the Directory contains between the scientific information and the people involved in its production and usage. Because of that, the reach of this public content is highly desirable.

In particular, reachability via web search engines can have a high impact over the lifespan of the project knowledge and its spread. Because of its curated content and its interrelation with internal and external resources, the Directory has a potential for enhanced visibility in search engines. We have built on this potential by enriching some of the web content through RDFa, a mark-up language that makes the types, relationships and attributes in the Directory understandable by other software, and in particular by search engines such as Yahoo or Google [50, 51].

We have monitored traffic data since the Directory went public in January 2010 (Figure 3). Traffic initially decreased towards the end of the project but thereafter started to increase by a steady 10% on a month-on-month basis, with the vast majority of traffic originating from web searches. This pattern suggests a shift in the usage of DC-RESEARCH.EU from a project-specific resource to the wider external public, which shows that the information generated within the network is still 'alive' after the end of the project.

An inspection of the most used search keywords reflects the content of the Directory: people, resources and the combination of the two without revealing any particular 'artifact' (see also Figure 5 in the Discussion section).

Integrability

The use of Semantic Web technologies also allows the Directory to integrate external tools and functionalities at minimal cost. We show this point by means of RelFinder [52], a tool that was originally developed in the context of the DBPedia project [53] for analysis and visualization of entities represented in a Semantic Web-enabled knowledge base.

RelFinder asks the user for two or three entities and, after a disambiguation step, searches for relevant connections in the knowledge base that connect such terms and displays the result as a graph. This functionality fits well within the Directory, allowing, for instance, the discovery of connections amongst participants, or between a given researcher and a specific data set. An example of a result found via RelFinder is presented in Figure 4.

Integration of RelFinder in the Directory is a significant functional enrichment, which required only the configuration of the SPARQL end-point of the Directory: a single web address.

Portability and long-term persistence

Conversely, adoption of Semantic Web technologies allows the Directory content to be readily integrated with external resources. As an extreme case of this integrability, we have migrated the entire contents of the Directory to the Talis platform [54]: a public infrastructure offering reliable and efficient storage and access of both unstructured and structured data. (Talis, which is behind the publication of UK government data online [55], offers free access to its platform to qualifying academic projects under the terms of the Talis Connected Commons program [56]).

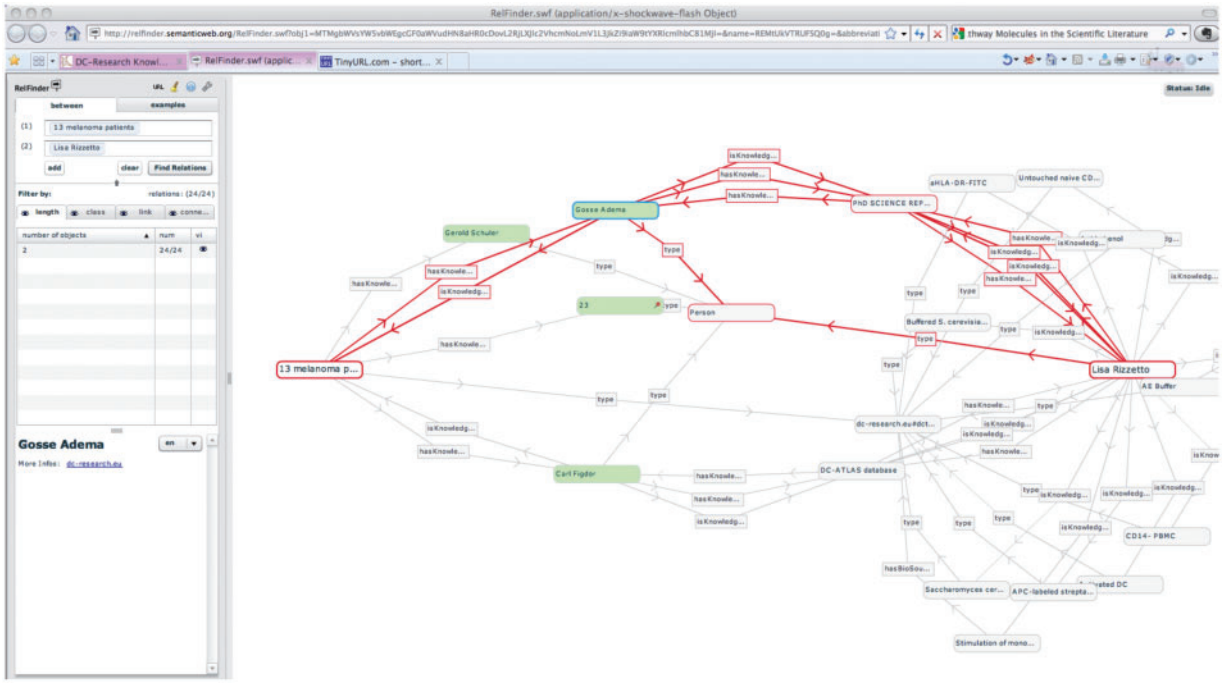

Figure 4: An example of the use of RelFinder to find relationships, in the Directory, between a researcher and a given data set. This search modality goes consistently beyond a single text search, and allows one to find a 'contact point' for a resource of interest. The results reported in the figure can be reproduced by accessing the URL http://tiny.cc/dcdrfdemo.

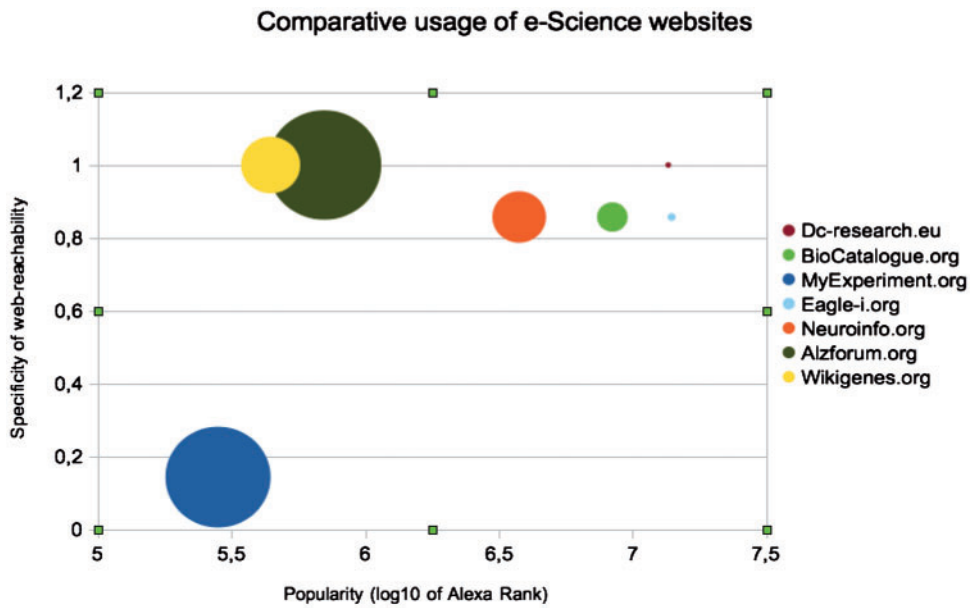

Figure 5: Comparison of web-traffic of the DC-THERA Directory and other related resources. The image reports traffic information for some of the information resources presented in the discussion section, as collected from the alexa.com web traffic monitoring service over the period April to June 2011. The *x*-axis reports the Alexa rank, which is a measure of web-traffic. Numbers are log scaled and range from 703 100 (most visited) for the Alzheimer research forum to 14 009 573 (least viewed) for Eagle-i. The *y*-axis reports the percentage of the top seven search queries that are relevant to the content of the website. Terms not evidently pertaining to the content of the website have been double checked with Google queries for the term, with scope limited to the website domain. If in doubt, terms have been considered pertinent. 'Pertinence' is not related to the performance of individual sites, but rather to the specificity with which a generic query on the web can reach them. Finally, the size of the dots indicated the number of resources linking to the corresponding web resource. All measures from http://alexa.com are derived from a panel of users of which the suitability for the purpose of this study cannot be assessed. These measures should only be considered as indicative.

Tools to access and query the content of the Directory (e.g. RelFinder) can be directed seamlessly to the Directory SPARQL end-point, or to the end-point that is provided by the Talis platform [57].

In this way we have achieved two important results. First, we have guaranteed maintenance of the Directory content even beyond the availability of funds to operate the current web infrastructure.

Second, we have accessed a range of functionalities offered by the Talis platform, including access to new dissemination channels such as those being explored on the data market platform, Kasabi [58], which currently offers the content of the Directory to interested early adopters.

TECHNICAL NOTES

Semantic web

The DC-THERA Directory makes extensive use of the Semantic Web framework, which is a set of standards and technologies designed to make the web a distributed, query-able, knowledge base. Annotation in the Directory closely matches the data model defined by the Resource Description Framework (RDF) [59], a key component of this framework. The Directory provides information on the web through different Semantic Web technologies, such as RDFa and SPARQL, mentioned above. More precisely, each resource in the Directory is associated with a URI, which is also an URL (e.g. 'http://dc-research.eu/rdf/protocol/10') pointing to an RDF representation of resource information (serialized in XML/RDF). A related URI/URL (e.g. 'http://dc-research.eu/protocol/10') resolves to an HTML representation of the information, which is enriched via an RDFa mark-up. Interaction with the information content in the Directory is based on the REST paradigm [60].

Public information presented by the Directory can also be queried by means of SPARQL, a query language for RDF-based knowledge bases. A SPARQL end-point (i.e. a server that can answer SPARQL queries) is available at the address http://dc-research.eu/sparql.

Software infrastructure

Development of the Directory within the lifecycle of the project has required rapid prototyping and agile methodologies, as discussed in [61]. The Directory has been developed in cycles of releases (Table 2). At the end of each cycle, feedback on functionalities and prioritization of the next functionalities to be implemented has been collected from its end-users: the network participants.

The Directory is based on an ad-hoc software engine that combines object oriented modeling of the main types in the Directory with a schema-less RDF-like modeling of information, following a similar approach to that presented in [62]. This engine was developed to enable the usage of established web development techniques and frameworks and thus to maximize the maintainability of the code base and the effectiveness of the deployment cycle. In particular, the design of the Directory follows a Model-View-Controller (MVC) approach [63], implemented through the Symfony [64] framework [63] and the Relational mapper (ORM) engine Doctrine [65]. Implementation of the Semantic Web functionalities has been based largely on the ARC Library [66].

Integration of external resources

The Directory integrates a range of features from other computational resources, both public ones and resources where access is restricted to DC-THERA participants.

ArrayExpress Atlas [67] and Whatizit [68] are accessed through publicly available web services to provide information on public data sets and external public literature repositories, respectively.

DC-THERA-specific databases and services such as DC-BASE [69] and Pathway Analysis services [70] are accessed through ad-hoc developed REST-based interfaces.

Other resources are imported into the Directory through specialized scripts, such as BioLexicon [71], which is used to expand terms in their synonym and morphological variants, and the body of ontologies, expressed in OWL, which constitute the DC-THERA Ontology.

DISCUSSION

The DC-THERA Directory addresses data management issues typical of a large collaborative biomedical research effort, and in particular the need for the information produced to be part of a larger shared information space. From a wider perspective, the Directory is part of a range of modern developments that affect the way science is conducted and communicated.

Table 2: DC-THERA project history and user feedback

Release	Features
Oct 2009	Search/Browse functionality
	Main contents
Feb 2009	Backend with editing functions, available to selected users
	Contents from all DC-THERA Scientific reports included by curators
Jul 2009	Editing back-end available to all users
	External services integrated (e.g. *ArrayExpress, Pathway Processor, Links from persons to PUBMED articles, WhatIzIt used for Text Tagging with ontologies*)
Autumn 2009	*Standard Operating Procedures (SOP) added as protocols*
	Graphical look improved
	Tooltips for categories and acronyms added
Mid 2010	RDF/RDFa/SPARQL export
	Relfinder integration
	Links between protocols and bio-materials used were added
	Contents and their classification reviewed
Mid 2011	Content updates
	RDF dump loaded in Talis

The table summarizes the features introduced in the Directory over time. The Directory development followed an iterative approach and at each release user feedback was gathered to plan and prioritize next developments. Reported in italic are the features requested by users, and not originally planned by the steering committee.

The commitment of the Directory to the web and Semantic Web standards reflects the increasing role of the web as a knowledge mediation platform [7] and the emergence of standard publication practices such as Linked Data [72]. The attention that the Directory pays to both the social aspect of annotation and its machine readability reflects the trend towards the development of communities of interest [73] and towards the formalization of the research process, which is explored in [74]. The cell-type specific annotation of resources within the Directory provides a significant improvement in the way researcher can access, share and relate information.

Related work

Several other resources are pioneering the development of computational collaborative research tools to support research in the Life Sciences. We provide a brief review here of some representative examples, rather than an exhaustive list.

Some social e-science sites have been developed with a specific need or data type as a unifying item on which a community was later built. This is the case of myExperiment [75], a social site designed for the exchange of bioinformatic workflows. It supports annotation via RDF and ontologies and publishes information via SPARQL in a similar way to the Directory. Similar features are offered by BioCatalogue [76] for the annotation of bioinformatic web services. Within the Systems Biology project SysMO, SymoDB [77] has been developed to support sharing of models and simulations among participants.

Other sites are intended as gateways for specific research communities. Similar to the Directory, they aggregate and organize heterogeneous information, but they vary in the specifics of implemented solutions, and include the following.

The Neuroscience Information Framework [78] is a comprehensive web resource for the Neuroscience field that makes several types of ontology-annotated resources available, providing data federation for many different biological databases and advanced search features.

The Alzheimer Research Forum [79] collects information about Alzheimer disease in a similar way. It allows users to link resources to scientific hypotheses and to discussions about them.

The Trial Item Manager [80] is an application similar to the Directory, allowing for collaborative editing of clinical trial information by means of detailed case report forms.

Particularly similar to the Directory and to its design as an 'actionable' inventory of research assets

is Eagle-I [81], a recently formed consortium of several US organizations, aiming to, in their words: 'build a prototype of a national research resource discovery network—one that will help biomedical scientists search for and find previously invisible, but highly valuable, resources'. Mentioned examples of these resources are: animal models, reagents, cell and tissue banks, core facilities and training opportunities.

A different class of information systems to support the collaborative development of information resources are wikis, which are the backbone of many research project information systems as well as of large-scale annotation efforts [82–84]. However, while wikis are an effective tool in many cases, they have limitations in scaling up with structured non-regular data [85]. Tools such as the Directory are designed to address information with these characteristics.

We have attempted a comparison of the usage of these resources with that of the DC-THERA Directory (Figure 5). While the numbers reported can only be considered as indicative, they show that the Directory is substantially less frequented than information resources that have an established web presence (as hinted by the number of incoming links) and which appeal to a relatively generic public. This is not surprising, as the Directory is both a new and specialized resource. However, the specificity of queries is among the highest of the resources presented (all top queries that lead to the website are relevant for its content). Together with results reported in Figure 3, this indicates a healthy status for the Directory, which has evolved from a project-specific information resource into a web resource that is attracting (and retaining) new users clearly focused on its content.

Limitations and perspectives

There remain limitations in the adoption of collaborative web environments in the research practice that vary depending on the characteristics of the project and its social environment.

In the case of the Directory, there was no problem of creating a community, since a research network was already in place at the time of its design. In our experience, the bottleneck in the uptake of this environment was the engagement of users to provide information for the Directory, which was solved partly via automated information importers and curation. By means of user experience surveys and feedback collected at demonstrations, we found that the Directory had a good reception among participants. Beside qualitative observations, we organized polls from a representative panel of selected participants, and the Directory was rated high (>7/10) on aspects such as the overview it provides, its intuitiveness and its search functionalities. Users' feedback was also important to reveal limitations of the Directory, and it lead to the introduction of new features and improvements, as highlighted in Table 2. The use of the Directory through tools such as RelFinder has been of particular interest as it provides an innovative and intuitive way to mine connections among participants and knowledge. However its usage still requires the mediation of an expert, as the low level representation of information presented by RDF and its mix of domain and 'meta' statements can be confusing to a biomedical research public. Even more promising, though, is the increasing number of visits that the Directory is collecting from Web Searches (Figure 3).

Overall, the main issue in the development of social resources for science is rewarding content provision. 'Web visibility' could be a reward that we are exploring in the Directory. In the past, funding agencies and scientific journals had a key role in consolidating the role of databases and standards in the research community [86]. Similar incentives could benefit the development of coherent data management strategies in the research practice. Another interesting incentive could be linking knowledge management systems to project administration, for instance, by automating project reporting.

A distinct problem relates to the complexity of curation, for which there is not an easy solution. In general, there is a trade-off between coverage and precision of annotation. In the Directory we have resorted to curated ontological information, which would have been difficult to crowd-source, at this stage of the evolution of technology.

CONCLUSIONS

The DC-THERA Directory has explored the use of a Semantic Web-based data management platform for the curation of the research assets, or the 'network know-how' of a research network.

Within the DC-THERA community, the Directory has proved important in stimulating data

integration and collaborative research by sharing information.

Controlled vocabularies facilitated data integration in immunology as well as comparison of large 'omics' data sets by annotation of cell type specific processes and variables. Emphasis on ontological annotations and standards makes it a resource valuable beyond the limits of the project and, in particular, we have shown how the Directory can support important aspects of the data management lifecycle, providing a resource-efficient way to integrate the information content with external resources, such as tools and knowledge bases.

Key points

- The collaborative and computational nature of 'big data' research requires the development of knowledge-management solutions, based on shared and machine-processable annotations.
- The DC-THERA Directory is a web resource and a collaborative platform for translational immunology focused on the activities and expertise of a multi-national research project.
- The information management of collaborative research projects can be improved by the adoption of ontologies and standard representations to maximize visibility, reachability and maintainability of research information during and beyond the project lifecycle.
- Compliance with standards offers an economic advantage by allowing resource-effective integration of third-party tools and enabling the use of public repositories for unstructured data and the use of data-economy platforms.
- Cell-type specific annotation of research resources in immunology can rely on ontologies to enhance data integration, sharing and collaboration among researchers.

Acknowledgements

The authors would like to thank all the DC-THERA participants, in particular those who participated in the definition of the application requirements and who took part in the evaluation panel, Maria Cristina Gauzzi, Damariz Rivero, Éva Rajnavölgyi and Rita Nunes. They wish to thank Olivier Lefevre for the work on the DC-BASE integration and Luca Beltrame for his work in the integration of Pathway Analysis functionalities.

FUNDING

This work was supported by the DC-THERA Network of Excellence (European Commission NoE contract number: LSHB-CT-2004-512074).

References

1. Szalay A, Gray J. 2020 computing: science in an exponential world. *Nature* 2006;**440**:413–4.
2. Antezana E, Kuiper M, Mironov V. Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief Bioinform* 2009;**10**:392–407.
3. Rebholz-Schuhmann D, Nenadic G. Biomedical semantics: the hub for biomedical research 2.0. *J Biomed Semantics* 2010;**1**:1.
4. Burks C, Fickett J, Goad W, *et al*. The GenBank nucleic acid sequence database. *Comput Appl Biosci* 1985;**1**:225–33.
5. Brazma A, Parkinson H, Sarkans U, *et al*. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003;**31**:68–71.
6. Barrett T, Suzek T, Troup D, *et al*. NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res* 2005;**33**:D562–6.
7. Goto S, Bono H, Ogata H, *et al*. Organizing and computing metabolic pathway data in terms of binary relations. *Pac Symp Biocomput* 1997;175–86.
8. Roberts RJ. PubMed Central: the GenBank of the published literature. *Proc Natl Acad Sci USA* 2001;**98**:381–2.
9. Gardiner-Garden M, Littlejohn T. A comparison of microarray databases. *Brief Bioinform* 2001;**2**:143–58.
10. Brazma A. On the importance of standardisation in life sciences. *Bioinformatics* 2001;**17**:113–4.
11. Brinkman R, Courtot M, Derom D, *et al*. Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 2010;**1(Suppl 1)**:S7.
12. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform* 2008;**41**:687–93.
13. Stein L. Integrating biological databases. *Nat Rev Genet* 2003;**4**:337–45.
14. Stein L. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat Rev Genet* 2008;**9**:678–88.
15. Etzold T, Argos P. SRS–an indexing and retrieval tool for flat file data libraries. *Comput Appl Biosci* 1993;**9**:49–57.
16. Ruttenberg A, Clark T, Bug W, *et al*. Advancing translational research with the Semantic Web. *BMC Bioinformatics* 2007;**8(Suppl 3)**:S2.
17. Hu J, Mungall C, Nicholson D, *et al*. Design and implementation of a CORBA-based genome mapping system prototype. *Bioinformatics* 1998;**14**:112–20.
18. Wilkinson MD, Links M. BioMOBY: an open source biological web services proposal. *Brief Bioinform* 2002;**3**:331–41.
19. Hull D, Wolstencroft K, Stevens R, *et al*. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 2006;**34**:W729–32.
20. Wilkinson MD, McCarthy L, Vandervalk B, *et al*. SADI, SHARE, and the in silico scientific method. *BMC Bioinformatics* 2010;**11(Suppl 12)**:S7.
21. Hucka M, Finney A, Sauro H, *et al*. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;**19**:524–31.
22. Demir E, Cary M, Paley S, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010;**28**:935–42.
23. Ashburner M, Ball C, Blake J, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
24. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–5.

25. Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform* 2004;**102**:20–38.
26. Bechofer S, van Harmelen F, Hendler J, *et al*. OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/ (15 April 2011, date last accessed).
27. W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview. http://www.w3.org/TR/owl2-overview/ (15 April 2011, date last accessed).
28. Tirmizi SH, Aitken S, Moreira DA, *et al*. Mapping between the OBO and OWL ontology languages. *J Biomed Semantics* 2011;**2(Suppl 1)**:S3.
29. Hoehndorf R, Oellrich A, Dumontier M, *et al*. Relations as patterns: bridging the gap between OBO and OWL. *BMC Bioinformatics* 2010;**11**:441.
30. Rubin DL, Lewis SE, Mungall CJ, *et al*. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS* 2006;**10**:185–98.
31. Noy N, Shah N, Whetzel P, *et al*. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;**37**:W170–3.
32. Bard JBL, Rhee SY. Ontologies in biology: design, applications and future challenges. *Nat Rev Genet* 2004;**5**:213–22.
33. Bourne P. Will a biological database be different from a biological journal? *PLoS Comput Biol* 2005;**1**:179–81.
34. Howe D, Costanzo M, Fey P, *et al*. Big data: The future of biocuration. *Nature* 2008;**455**:47–50.
35. Shotton D, Portwin K, Klyne G, *et al*. Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS Comput Biol* 2009;**5**:e1000361.
36. Clark T. Knowledge integration in biomedicine: technology and community. *Brief Bioinform* 2007;**8**:E1–3.
37. Silvertown J. A new dawn for citizen science. *Trends Ecol Evol* 2009;**24**:467–71.
38. Neylon C, Wu S. Open Science: tools, approaches, and implications. *Pac Symp Biocomput* 2009;**2009**:540–4.
39. Splendiani A, Brandizi M, Even G, *et al*. The genopolis microarray database. *BMC Bioinformatics* 2007;**8(Suppl 1)**:S21.
40. Data's shameful neglect. *Nature* 2009;**461**:145.
41. Mons B, Haagen H, van, Chichester C, *et al*. The value of data. *Nat Genet* 2011;**43**:281–3.
42. DC-THERA European Network — DC-THERA. http://www.dc-thera.org/ (6 April 2011, date last accessed).
43. Brandizi M, Guendel M, Scognamiglio C, *et al*. DC-THERA Directory, a Knowledge Management System for the support of the European Dendritic Cell Immunology Community. In: *Proceedings of Semantic Web Application and Tools for Life Sciences,* Amsterdam, 2009. Vol-559: Demo2.pdf. CEUR-WS.
44. Guendel M, Scognamiglio C, Brandizi M, *et al*. The DC-THERA Directory: A Knowledge Management System to Support Collaboration on Dendritic Cell and Immunology Research. *NETTAB 2009 Conference Proceedings*. Genova: Edizioni Libro di Scrivere, 2009.
45. DC-Research Knowledge Portal. http://dc-research.eu/ (6 April 2011, date last accessed).
46. Cruz IF, Gjomemo R, Jarzab G. An interoperation framework for secure collaboration among organizations. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*. San Jose, CA, USA: ACM, 2010.
47. Berners-Lee T, Hendler J, Lassila O. The Semantic Web. *Scientific American* 2001;34–43.
48. RDFa Primer. http://www.w3.org/TR/xhtml-rdfa-primer/ (12 April 2011, date last accessed).
49. SPARQL Protocol for RDF. http://www.w3.org/TR/2005/WD-rdf-sparql-protocol-20050527/ (12 April 2011, date last accessed).
50. RDF and the Monkey YDN Blog. http://developer.yahoo.com/blogs/ydn/posts/2008/05/rdf_and_the_mon/ (5 May 2011, date last accessed).
51. Official Google Webmaster Central Blog: Introducing Rich Snippets. http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html (12 April 2011, date last accessed).
52. Lohmann S, Heim P, Stegemann T, *et al*. The RelFinder user interface: interactive exploration of relationships between objects of interest. In: *Proceedings of the 15th international conference on Intelligent user interfaces*. Hong Kong, China: ACM, 2010;421–2.
53. Bizer C, Lehmann J, Kobilarov G, *et al*. DBpedia - A crystallization point for the Web of Data. *J. Web Semantics* 2009; **7**:154–65.
54. Talis Platform - Home. http://www.talis.com/platform/ (13 April 2011, date last accessed).
55. data.gov.uk | Opening up government. http://data.gov.uk/ (13 April 2011, date last accessed).
56. Talis Platform - Connected Commons. http://www.talis.com/platform/cc/ (13 April 2011, date last accessed).
57. dcresearch provided by the Talis Platform.
58. Kasabi. http://www.kasabi.com/ (13 April 2011, date last accessed).
59. RDF Primer. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/ (13 April 2011, date last accessed).
60. Fielding R, Software D, Taylor R. Principled design of the modern web architecture. *ACM Trans Internet Techn* 2002;**2**: 115–20.
61. De Roure D, Goble C. Software design for empowering scientists. *Software. IEEE* 2009;**26**:88–95.
62. Puleston C, Parsia B, Cunningham J, *et al*. Integrating object-oriented and ontological representations: a case study in Java and OWL. In: Proceedings of the International Semantic Web Conference (ISWC), Karlsruhe, Germany, 2008. LNCS 5318:130-45. Springer.
63. Gamma E, Helm R, Johnson R, *et al*. *Design Patterns: Elements of Reusable Object-Oriented Software*. Boston: Addison Wesley, 1995.
64. Symfony - Web PHP Framework. http://www.symfony-project.org/ (5 May 2011, date last accessed).
65. Doctrine - PHP Object Persistence Libraries and More. http://www.doctrine-project.org/ (5 May 2011, date last accessed).
66. Home - GitHub. https://github.com/semsol/arc2/wiki (5 May 2011, date last accessed).
67. Kapushesky M, Emam I, Holloway E, *et al*. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res* 2010;**38**:D690–8.

68. Rebholz-Schuhmann D, Arregui M, Gaudan S, *et al*. Text processing through Web services: calling Whatizit. *Bioinformatics* 2008;**24**:296–8.
69. DC-BASE. http://dc-base.dc-atlas.net/ (5 May 2011, date last accessed).
70. Cavalieri D, Castagnini C, Toti S, *et al*. Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases. *Bioinformatics* 2007;**23**:2631–2.
71. Sasaki Y, Montemagni S, Pezik P, *et al*. Biolexicon: A lexical resource for the biology domain. *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM)* 2008. Turku, Finland, 109–16.
72. Bizer C, Heath T. Linked data-the story so far. *IJSWIS* 2009;**5**:1–22.
73. Webster Y, Dow E, Koehler J, *et al*. Leveraging health social networking communities in translational research. *J Biomed Informatics* 2011;**44**:536–44.
74. Ciccarese P, Wu E, Wong G, *et al*. The SWAN biomedical discourse ontology. *J Biomed Informatics* 2008;**41**:739–51.
75. Goble C, Bhagat J, Aleksejevs S, *et al*. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res* 2010;**38**:W677–82.
76. Bhagat J, Tanoh F, Nzuobontane E, *et al*. BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res* 2010;**38**:W689–94.
77. SysMO-DB. http://www.sysmo-db.org/ (5 May 2011, date last accessed).
78. Gardner D, Akil H, Ascoli G, *et al*. The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics* 2008;**6**:149–60.
79. Kinoshita J, Clark T. Alzforum. *Methods Mol Biol* 2007;**401**: 365–81.
80. Mucke R, Lobe M, Knuth M, *et al*. A semantic model for representing items in clinical trials. *2009 22nd IEEE International Symposium on Computer-Based Medical Systems,* 2009. Albuquerque, NM, USA: ;1–8.
81. The eagle-i consortium. eagle-i home page. https://www.eagle-i.org/home/ (11 May 2011, date last accessed).
82. Mons B, Ashburner M, Chichester C, *et al*. Calling on a million minds for community annotation in WikiProteins. *Genome Biol* 2008;**9**:R89.
83. Huss J, Lindenbaum P, Martone M, *et al*. The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res* 2010;**38**:D633–9.
84. Waldrop M. Big data: Wikiomics. *Nature* 2008;**455**:225.
85. Arita M. A pitfall of wiki solution for biological databases. *Brief Bioinform* 2009;**10**:295–6.
86. Ball CA, Brazma A, Causton H, *et al*. Submission of microarray data to public repositories. *PLoS Biol* 2004;**2**:E317.
87. Masci A, Arighi C, Diehl A, *et al*. An improved ontological representation of dendritic cells as a paradigm for all cell types. *BMC Bioinformatics* 2009;**10**:70.
88. Bard J, Rhee S, Ashburner M. An ontology for cell types. *Genome Biol* 2005;**6**:R21.
89. Malone J, Holloway E, Adamusiak T, *et al*. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 2010;**26**:1112–8.
90. Whetzel P, Parkinson H, Causton H, *et al*. The MGED ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics* 2006;**22**:866–73.
91. de Matos P, Alcántara R, Dekker A, *et al*. Chemical entities of biological interest: an update. *Nucleic Acids Res* 2010;**38**: D249–54.
92. Sayers E, Barrett T, Benson D, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2011;**39**:D38–51.