publications

Article

A Vision for Open Cyber-Scholarly Infrastructures

Costantino Thanos

Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR), Via G. Moruzzi 1, Pisa 56124, Italy; Costantino.Thanos@isti.cnr.it; Tel.: +39-050-315-2910; Fax: +39-050-315-3464

Academic Editor: Craig Smith
Received: 22 February 2016; Accepted: 16 May 2016; Published: 24 May 2016

Abstract: The characteristics of modern science, *i.e.*, data-intensive, multidisciplinary, open, and heavily dependent on Internet technologies, entail the creation of a linked scholarly record that is online and open. Instrumental in making this vision happen is the development of the next generation of Open Cyber-Scholarly Infrastructures (OCIs), *i.e.*, enablers of an open, evolvable, and extensible scholarly ecosystem. The paper delineates the evolving scenario of the modern scholarly record and describes the functionality of future OCIs as well as the radical changes in scholarly practices including new reading, learning, and information-seeking practices enabled by OCIs.

Keywords: scholarly record; linked scholarly record; semantic publishing; enhanced publication; linked data; scientific article models; information exploration; topic map; reading practices; learning practices; information seeking

1. Introduction

Modern science has undergone deep transformations due to recent advances in information technology, computer infrastructures, and the Internet as well as the development of new high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, and sensor networks that are generating huge volumes of research data. Modern science is increasingly based on data-intensive computing; it tries to solve complex problems not within a discipline but across disciplines (multidisciplinary/interdisciplinary science); and it is conducted by scientists at different locations at the same (synchronous) or different (asynchronous) times by collapsing the barrier of distance and removing geographic location as an issue. Finally, there is an emerging consensus among the members of the academic research community that the practices of modern science should be congruent with "Open Science". Global scientific collaboration takes many forms, but from the various initiatives around the world a consensus is emerging that collaboration should aim to be "open" or at least should include a substantial measure of "open access" to the results of research activities.

This new science paradigm and a revolutionary process of digitization of information have created enormous pressure for radical changes in scholarly practices. They have induced changing demands and created expectations of scholars that are significantly different than they were just a few years ago. Today one of the main challenges faced by scholars is to make the best use of the world's growing wealth of scientific information. Scholars need to be able to find the most authoritative, comprehensive, and up-to-date information about an important topic; to find an introduction to a topic that is organized by an expert; to conduct perspective analyses of scientific literature (for example, what arguments are there to refute this article?); to conduct a lineage analysis (for example, where did this idea come from?); *etc.* They also need to be able to navigate through an information-rich environment in order to discover useful knowledge from data, *i.e.*, to extract high-level knowledge from low-level data in the context of huge volume datasets.

A further pressure for radical changes in scholarly practice is caused by the revolutionary changes underway in scientific communication. These include:

- Scientific data are becoming key to scientific communication; as such they must be integrated with scientific publications in order to support repeatability, reproducibility, and re-analyses.
- Scientific data and publications have to cross disciplinary boundaries; therefore, in order to maintain the interpretative context they must be semantically enhanced, *i.e.*, semantic mark-up of textual terms with links to ontologies/terminologies/vocabularies, interactive figures, *etc.* Semantic services will help readers to find actionable data, interpret information, and extract knowledge.
- Scientific literature is becoming increasingly online. The digital form of the article has permitted the definition of new scientific article models, based on modularization techniques, which allow the overcoming of the traditional linear form of the scientific article.
- Advanced linking technology allows the meaningfully interconnection of datasets and article modules in many different ways. This permits the creation of several patterns of interest for scholars and scientists.

Instrumental in making radical changes in scholarly practices happen is the development of the next generation of Open Cyber-scholarly Infrastructures (OCIs). In a previous paper [1] I have delineated the future of the digital scholarship and argued that connectivity is its technological foundation. In this paper I have further elaborated this concept and argue that building future OCIs will contribute to radical changes in the way scholars create, communicate, search for, and consume scientific information. OCIs have the potential to completely reshape scientific research.

The paper is organized as follows: in Section 2 the evolving scenario of the modern scholarly record is described. In Section 3 a linked scholarly record that meets the needs of modern science is described. Section 4 describes the radical changes in the scholarly practices enabled by connectivity and semantic technologies. Section 5 describes the functionality of the future OCIs. Finally, Section 6 contains some concluding remarks.

2. The Modern Scholarly Record

A scholarly record is taken as a means of aggregation of scientific journals, gray literature, and conference presentations plus the underlying datasets and other evidence to support the published findings. Moreover, the communications of today's scholars encompass not only journal publications and underlying datasets but also less formal textual annotations and a variety of other work products, many of them made possible by recent advances in information technology and Internet. This evolving scholarly record can also include news articles, blog posts, tweets, video presentations, artworks, patents, computer code, and other artifacts.

This record is highly distributed across a range of libraries, institutional archives, publishers' archives, discipline-specific data centers, and institutional repositories. It is also poorly connected, and this constitutes a major obstacle to full engagement by scholars.

Two of the main constituents of the modern scholarly record are the *scientific dataset* and the *scientific article*.

3. The Scientific Dataset

There is no single well-defined concept of *dataset*. Informally speaking, we can think of a dataset as a meaningful collection of data that is published and maintained by a single provider, deals with a certain topic, and originates from a certain experiment/observation/process. In the context of the Linked Data world, a dataset means a set of RDF triples that is published, maintained, or aggregated by a single provider.

The concept of collection also suggests that there is an *intentional collecting* of the constituents of a dataset.

In [2] different kinds of relatedness among the grouped data have been identified:

Circumstantial Relatedness: a dataset is thought of as consisting of data related by time, place, instrument, or object of observation.

Syntactic Relatedness: Data in a dataset are typically expected to have the same syntactic structure (records of the same length, field values in the same places, *etc.*).

Semantic Relatedness: Data in a dataset may be about the same subject or make assertions similar in content.

A dataset, once accepted for deposit and archived, is assigned by a Registration Agency a *Digital Object Identifier* (DOI) for registration. A Digital Object Identifier (DOI) is a unique name (not a location) within a name space of a networked data environment and provides a system for persistent and actionable identification of datasets. It must: (i) unambiguously identify the dataset; (ii) be globally unique; and (iii) be associated with a naming resolution service that takes the name as input and shows how to find one or more copies of the identical dataset.

A dataset must be accompanied by *metadata*, which describes the information contained in the dataset, details of data formatting and coding, how the dataset was collected and obtained, associated publications, and other research information. Metadata formats range from a text "readme" file, to elaborate written documentation, to systematic computer-readable definitions based on common standards.

The DOI as a long-term linking option from data to source publication is of fundamental importance.

DOIs could logically be assigned to every single data point in a dataset; in practice, the assignment of a DOI is more likely to be to a meaningful set of data following the index Principle of Functional Granularity*: identifiers should be assigned at the level of granularity appropriate for the functional use that is envisaged*.

However, having the ability to make references to subsets of datasets would be highly desirable. Datasets may be subdivided by row, by column, or both.

Devising a simple standard for describing the chain of evidence from the dataset to the subset would be highly valuable. The task of creating subsets is relatively easy and is done in a large variety of ways by researchers.

With respect to the *versioning problem*, *i.e.*, how to treat subsequent versions of the same dataset, it is recommended to treat them as separate datasets.

An emerging "best practice" in the scientific method is the process of publishing scientific datasets. Dataset Publication is a process that allows the research community to discover, understand, and make assertions about the trustworthiness and fitness for purpose of the dataset. In addition, it should allow those who create datasets to receive academic credit for their work. The ultimate aim of Dataset Publication is to make scientific datasets available for reuse both within the original disciplines and the wider community.

The Dataset Publication process is composed of a number of procedures that altogether implement the overall functionality of this process. In particular, they should support the following functionality relevant for achieving dataset reusability: (i) dataset peer-reviewing; (ii) dataset discoverability; (iii) dataset understandability; and (iv) making dataset assessable.

4. The Modern Scientific Article

We foresee that in an increasingly online and interconnected scientific world the structure, functionality, and presentation of the scientific article is destined to change radically. The article will become a window for scientists and scholars, allowing them to not only actively understand a scientific result, but also to reproduce it or extend it; it will act as an access point for, or interface to, any type of global networked resource. Another way of viewing the modern article is as an interface through which authors and readers interact.

The future digital scientific article will feature several important characteristics:

First, modularization or disaggregation of the scientific article, *i.e.*, the linear form of the scientific article will be overcome and it will be presented as a network of modules meaningfully connected by relations. In essence, a new modular model of scientific article will emerge with two main constituents: modules and relations. The modules are conceptual information units representing self-contained, though related types of information. They can contain *organizational information* concerning the structural aspects of an article as well as *scientific discourse information* concerning hypotheses made by the author of an article, evidence for the hypotheses, underlying datasets, findings, pointers to future research, *etc.* Modules could be located, retrieved, and consulted separately as well as in combination with related modules.

Different types of relations between the modules can be established: *organizational relations* that are based on the structure of the article; *discourse relations* that define the reasoning of the argument; *causal relations* that establish a causal connection between premise and conclusion; *comparison relations* where the relation is one of contradiction, similarity, or resemblance.

Deconstructing a scientific article into semantically typed modules will enable scientists and scholars to access and manipulate individual modules, such as hypotheses, conclusions, references, *etc.* It will allow a reader to compile her/his own version, depending on interests and background; the reader becomes a creator of her/his individual reading versions. In essence, modularization will allow a more flexible interaction between article author and reader.

Second, a digital scientific article is intrinsically dynamic, *i.e.*, mutable. This characteristic of the digital article allows the author to update it at any time, to revise it by changing its content, and expand it by adding annotations, hyperlinks, comments, *etc.* It also allows the inclusion of non-static information types such as animation, moving images, and sound.

Third, a digital scientific article can have embedded software that can allow one to, for example, compute a formula and visualize the results while reading the article; or to link to the underlying datasets, thus allowing the reader to perform additional analyses on the data. An example of the use of embedded software in articles is the concept of a multivalent document.

Several models for representing a scientific article have appeared in the literature. The name used to indicate these models is *enhanced publication*. Enhanced publication is a dynamic, versionable, identifiable compound of objects combining an electronic publication with embedded or remote research data, extra materials, post publication data, database records, and metadata. It is an umbrella concept that embraces many different article models.

The conceptual model of an enhanced publication includes a *mandatory text body* and a set of *interconnected sub-parts*. Several instantiations of this model have been proposed in the literature. These instantiations, essentially, regard the way the mandatory text body is organized, the type of the sub-parts, and the way they are connected to the text.

A first instantiation regards the case where the sub-parts are essentially *supplementary material* along with the mandatory text. Examples include presentation slides, appendixes to the text, tables, *etc.* In this case, generally, the sub-parts do not have an identifier and are not described by metadata.

A second instantiation regards the case where the mandatory text body is not a single block of text but is structured in a number of interconnected modules, such as abstract, sections, bibliography, *etc.*

A third instantiation regards the case where the sub-parts are scientific datasets external to the publication, *i.e.*, stored in discipline specific data centers/repositories with their own identity (DOIs). In this case, the scientific datasets are cited from within the text using a DOI system.

A fourth instantiation regards the case where some sections or modules of the text body or some sub-parts are *live*, meaning that they can be activated in order to produce visual content, video streaming, *etc.*

Finally, a fifth instantiation regards the case where some sections or modules of the text body or sub-parts can be dynamically executed at run time.

A generalization of the concept of "Enhanced Publication" is the concept of *Research Object* (RO). Informally, a Research Object is intended as a semantically rich aggregation of resources that poses some scientific intent or supports some research objective. It should allow a principled publication of the results of research activity in a self-contained manner that facilitates the sharing and reuse of these objects. An RO bundles together all the essential information relating to a scientific investigation, *i.e.*, article, data produced/used, methods used to produce and analyze that data, as well as the people involved in the investigation. In addition, an RO includes additional semantic information that allows one to link its components in a meaningful way.

Scientific articles are increasingly being assigned DOIs that provide live links from online citing articles to the cited articles in their reference lists. In addition, they should be enriched with appropriate metadata.

DOIs could logically be assigned to every single article module; having the possibility to make references to article modules would be highly desirable.

5. Linked Scholarly Record

The scholarly record is poorly interconnected. This is in opposition to modern science, which requires the establishment of discipline-specific linked scientific records in order to effectively support scholarly inquiry. In fact, scientists and scholars need to be able to move from hypotheses to evidence, from article to article, from dataset to dataset, and from article to dataset and conversely. They need to discover potentially significant patterns and ways to make meaningful connections between parts of the scholarly record.

From a conceptual point of view, a linked scholarly record means that its single parties, *i.e.*, a dataset, an article module, *etc.* constitute single nodes of a networked scholarly record that can be accessed by any scholar, anytime, anywhere.

The two pillars of the modern scholarly communication are discipline-specific Data Centers and Research Digital Libraries, whose technologies and organizations allow researchers to store, curate, discover, and reuse the data and publications they produce. Made to implement complementary phases of the scientific research and publication process, they are poorly integrated with one another and do not adopt the strengths of the other. Such a dichotomy hampers the realization of a linked scholarly record. However, I am confident that the recent technological advances in many fields of information technology will make it happen.

6. Linked Discipline-Specific Data Spaces

New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations are generating massive amounts of data. The availability of huge volumes of data is revolutionizing the way research is carried out and leading to a new data-centric way of thinking, organizing, and carrying out research activities. The most acute challenge stems from research teams relying on a large number of diverse and interrelated datasets but having no way to manage their scientific data spaces in a principled fashion.

An example taken from [3] illustrates the requirement for interlinking scientific datasets. "Consider a scientific research group working on environmental observation and forecasting. They may be monitoring a coastal ecosystem through weather stations, shore-and buoy-mounted sensors, and remote imagery. In addition, they can be running atmospheric and fluid-dynamics models that simulate past, current, and near-future conditions. The computations may require importing data and model outputs from other groups, such as river flows and ocean circulation forecasts. The observations and simulations are the inputs to programs that generate a wide range of data products, for use within the group and by others: comparison plots between observed and simulated data, images of surface-temperature distributions, animations of salt-water intrusion into an estuary. Such a group can easily amass millions of data products in just a few years. Soon, such groups will need to federate with other groups to create scientific data spaces of regional or national scope. They will need to easily export their data in standard scientific formats, and at granularities that do not necessarily correspond to the partitions they use to store the data."

Therefore, there is a need for mechanisms and approaches that allow the linking of datasets produced by diverse research teams. Linking a dataset refers to the capability of linking it to other external datasets, which in turn can be linked to from external datasets. Linking data will allow the sharing of scientific data on a global scale and interconnect data between different scientific sources. It also makes data access, *i.e.*, search and exploration, and data exploitation, *i.e.*, integration and reuse much easier. The process that enables the linking of datasets is known as *data publishing.*

A generalization of the linking data concept leads to the creation of linked scientific data spaces of disciplinary or interdisciplinary scope. A *data space* can be considered as an abstraction for the management of linked datasets. The concept of scientific data spaces responds to the rapidly-expanding demands of "data-everywhere".

A linked disciplinary-specific data space should enjoy the following properties:

- it contains datasets specific to a scientific discipline;
- any scientific community belonging to this discipline can publish on the scientific data space;
- dataset creators are not constrained by the choice of vocabularies with which to represent them;
- datasets are connected by links creating a global data graph that spans datasets and enables the discovery of new datasets;
- datasets are self-describing;
- datasets are strictly separated from formatting and presentational aspects;
- the scientific data space is open, meaning that applications do not have to be implemented against a fixed set of datasets, but can discover new datasets at run time by following the data links.

A managed linked data space will enable researchers to start browsing in one dataset and then navigate to related datasets; or it can support data search engines that crawl the data space by following links between datasets. However, in order to be able to implement a linked discipline-specific data space, the ability to meaningfully and formally describe the datasets that participate in the linked data space, as well as the links among them, is of paramount importance.

Metadata is the descriptive information about datasets that explains the measured attributes, their names, units, precision, accuracy, data layout, and ideally a great deal more. Most importantly, metadata should include the dataset lineage, *i.e.*, how the dataset was measured, acquired, or computed. Equally important is the concept of the dataset identifier, *i.e.*, DOI (or URI) as mechanisms for referring to datasets, on which there exists some agreement among multiple data providers.

Modeling the many kinds of relationships existing between datasets is equally important. We need to define metadata models for describing links. We must be able to model, for example, dataset B as a temporal/spatial abstraction of dataset A; or show that datasets A and B are generated independently but both reflect the same observational or experimental activity; or that datasets A and B were generated at the same time and by the same organization, *etc.*

In order to be able to exploit the full potential of the linked data space, it is importance to make sense of heterogeneous datasets that constitute a linked data space. This can be achieved by adopting formalisms for representing discipline-specific ontologies.

An initiative that implements the concept of linked data space by using the semantic Web technologies is Linked Data [4].

7. Linked Scientific Articles: Linked Literature Spaces

Above I have described the network-centric nature of the future scientific article. Deconstructing the scientific article into semantically typed modules allows the structuring of the scientific information as a multitude of interlinked modules. This will allow us to answer questions like: (i) what is the evidence for this claim? (ii) was this prediction accurate? (iii) what are the conceptual foundations for this idea? (iv) who has built on this idea? (v) who has challenged this idea, and using what kind of arguments? (vi) are there distinctive perspectives on this problem? and (vii) are there inconsistencies within this school of thought?

In essence, it will also enable the author to create paths of reasoning within the article as well as between articles. On the other hand, by following such paths the reader is enabled, for example, to assess the validity of a claim by gaining insight into its empirical backing.

In computational linguistics the structure and the relations of discourse has been extensively studied as well as the relationship between discourse semantics and information packaging (modularization). Some studies have suggested that the modularization of discourse is not based purely on semantics but that the rhetorical nature of discourse relations must also be taken into consideration when deconstructing a scientific article. It must also be pointed out that some segments of discourse play a subordinate role relative to previous segments they are connected to, while others are considered on a par; for example, the *Result* module has a coordinating role while the *Explanation* module is a subordinate one. This distinction, often called subordinating/coordinating, must also be considered when an article is decomposed into a number of modules.

In essence, breaking a scientific article into different modules is a difficult conceptual operation as it should take into consideration the discourse structure and relations.

In the literature many models of discourse relations have been proposed; as an example, a small set of eight relations has been proposed in order to support a principled modularization of a scientific article and a realistic scientific reasoning:

- Proves/refutes
- Supports/contradicts
- Agrees/disagrees
- Suggests/does not suggest

The discourse relations are materialized by explicitly labeled links. A link can be defined as a uniquely characterized, explicit, directed connection between modules that represents one or more different kinds of relations. We can have different types of links: *semantic links* implement relations of similarity, contrast, part of, *etc.*; *rhetorical links* implement relations of definition, explanation, illustration, *etc.*; and *pragmatic links* implement relations of prerequisite, usage, example, *etc.*

Modularization and linking will enable scientific information to become part of a global, universal, and explicit network of knowledge. Literature will be modeled as a network of modules. A generalization of the network centrality of scientific information leads to the creation of a **linked** scientific literature space of disciplinary or interdisciplinary scope. A scientific contribution thus becomes a rigorously connected, substantiated node or region in a linked scientific literature space.

However, in order to be able to implement a linked, discipline-specific literature space, it is important to meaningfully and formally describe the article modules that participate in the linked literature space as well as the links among them.

We need semantically rich metadata models to describe the article modules as well as the relations between them. Equally important is the concept of the module identifier, *i.e.*, DOI (or URI) as a mechanism for referring article modules on which there exist some agreements among multiple publishers.

8. Linking Literature Spaces with Data Spaces

The need to link datasets to scientific publications is starting to be held as a key practice, underpinning the recognition of data as a primary research output, rather than as a byproduct of research. Linking data to publications will enable scientists, while reading an article, to go off and look at the underlying data and even redo analyses in order to reproduce or verify results.

The distinction between data and publication is destined to disappear as both are made increasingly available in electronic form. It is the task of the linking technology to support the next step, *i.e.*, their integration.

Publishers are beginning to embrace the opportunity to integrate data with scientific articles but barriers to the sustainability of this practice include the sheer volume of data and the huge variety of data formats. Several levels of integration can be achieved ranging from tight to weak integration.

A tight integration is achieved when datasets are contained within peer-reviewed articles. In this publishing model, the publisher takes full responsibility for the publication of the article and the aggregated data embedded in it and the way it is presented. The embedding of the dataset into the publication makes it citable and retrievable. However, the reusability of the dataset is limited as it is difficult to find it separate from the publication. This publishing model is not appropriate when the embedded dataset is too large to fit into the traditional publication format. In addition, the preservation of these enhanced articles is more demanding than for traditional articles.

A less tight integration is achieved when the datasets reside in supplementary files added to the scientific article. The publisher offers authors the option of adding supplementary files to their article containing any relevant material that will not fit the traditional article format or its narrative, such as datasets, multimedia files, large tables, animations, *etc.* There are some issues related to this publishing model: they mainly concern the preservation of the supplementary files as well as the ability to find them independently from the main publication.

A weak integration is achieved when the datasets reside in Institutional Data Repositories or in discipline-specific Data Centers with bi-directional linking to and from articles. In this publishing model the article should include a citation and links to the dataset. The data preservation is the responsibility of the administrators of the Institutional Repository or Data Center. In this model the datasets become better discoverable and can be reused separately from the publication and in combination with other datasets. However, this publishing model depends very much on the existence of proper and persistent linking mechanisms enabling bi-directional citation. In the Big Data era it is obvious that only the weak integration scheme is viable. Unfortunately, due to technological and policy reasons discipline-specific Data Centers and Research Libraries currently do not interoperate.

Linking publications to the underlying data can produce significant benefits:

- help the data to be better discoverable
- help the data to be better interpretable
- provide the author with better credits for the data
- add depth to the article and facilitate better understanding.

Unifying all scientific datasets with all scientific literature to create a world in which data and literature interoperate, as in Jim Gray's vision, implies the capability to link Literature Spaces with Data Spaces, *i.e.*, the capability to create a *Linked Scholarly Record*.

Linking Literature Spaces with Data Spaces will increase scientific "information velocity" and will enhance scientific productivity as well as data availability, discoverability, interpretability, and reusability.

The main mechanism enabling the linking between datasets and articles in the scientific communication workflow is data citation. Data citation is the practice of providing a reference to datasets intended as a description of dataset properties that enable discover, interlinking, and access to the dataset. As such, proper citation mechanisms rely on the assignment of persistent identifiers to datasets, together with a description (metadata) of the dataset, which allows for discovery and, to some extent, reuse of the data. Several standards exist for citing datasets and practices vary across different disciplines and data repositories, supported by initiatives in various fields of applications.

9. Semantic Enhancement of the Scientific Record

A multidisciplinary approach to research problems draws from multiple disciplines in order to redefine a research problem outside of the usual boundaries and reach solutions based on a new understanding of complex situations. Scientific communication across disciplinary boundaries needs semantic enhancements in order to make the text intelligible to a broad audience composed of specialists in different scientific disciplines. This need motivated the current development of semantic publishing. By *semantic publishing* we mean the enhancement of the meaning of an online research article by automatically disambiguating and semantically defining specialist terms. This can be achieved by linking to discipline-specific ontologies and standard terminology repositories, by linking to other information sources of relevance to the article, and by direct linking to all of the article's cited references. Semantic mark-up of text is a technology that would facilitate increased understanding of the underlying meaning. Sophisticated text mining and natural language processing tools are currently being developed to recognize textual instances and link them automatically to domain-specific ontologies. Additional semantic enhancements can be obtained by intelligently linking scientific texts to third-party commentaries, archived talks, and websites.

Semantic publishing facilitates the automated discovery of an article, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers. It demands the enrichment of the article with appropriate metadata that are amenable to automated processing and analysis. The semantic enhancements increase the intrinsic value of scientific articles, by increasing the ease by which information, understanding, and knowledge can be extracted.

Semantic technologies are enabling technologies for semantic publishing.

In the context of multidisciplinary research, communities of research and data collections inhabit multiple contexts. There is the risk, when datasets are moving across contexts, of interpreting their representations in different ways caused by the loss of the interpretative context. This can lead to a phenomenon called "ontological drift" as the intended meaning becomes distorted when the datasets move across semantic boundaries (semantic distortion). This risk arises when a shared vocabulary and domain terminology are lacking.

Scientists nowadays face the problem of accessing existing large datasets by means of flexible mechanisms that are both powerful and efficient. Ontologies describe the domain of interest at a high level of abstraction and allow for expressing at the intentional level complex kinds of semantic conditions over such a domain. They are, thus, widely considered to be a suitable formal tool for sophisticated data access. Providing ontology-based access to data demands the creation of a conceptual view of data and presenting it to the scientist-user. This view is expressed in terms of an ontology and presents the unique access point for the interaction between the users and the system that manages the dataset.

The challenge is to link the ontology to a dataset that exists autonomously and has not been necessarily structured with the purpose of storing the ontology instances. In this case, the conceptual view and the datasets are at different levels of abstraction and are expressed in terms of different formalisms. For example, while logical languages are used to specify the ontology, datasets are usually expressed in terms of a data model. Therefore, there is a need for specific mechanisms for mapping the data to the elements of the ontology. In summary, in ontology-based data access, the mapping is the formal tool by which we determine how to link data to ontology, *i.e.*, how to reconstruct the semantics of datasets in terms of the ontology.

The main reason for a functionality that supports an ontology–based access to data is to provide high-level services to the scientists-clients. The most important service is query answering. Clients express their queries in terms of the conceptual view (the ontology) and the mapping and should translate the request into suitable queries posed to the system that manages the dataset.

In the context of a networked multidisciplinary scientific world, in order to maintain the interpretative context of data when crossing semantic boundaries, there is the need for aligning domain-specific ontologies that support the ontology-based access to distributed datasets. These ontologies are not standalone artifacts. They relate to each other in ways that can affect their meaning, and are distributed in a network of interlinked datasets, reflecting their dynamics, modularity, and contextual dependencies. Their alignment is crucial for effective and meaningful data access and usability. It is achieved through a set of mapping rules that specify a correspondence between various entities, such as objects, concepts, relations, and instances.

10. Innovation in Scholarly Practices

We are entering a new era characterized by the availability of huge collections of scientific articles. It is estimated that at least 114 million English-language scientific documents are accessible on the Web. Of these, it is estimated that at least 27 million are freely available.

Moreover, high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, and sensor networks are generating massive amounts of scientific data. This information explosion is making it increasingly difficult for scholars to meet their information needs. In addition, the availability of huge amounts of scientific information has caused scholars to significantly extend their search goals.

We expect that advanced semantic linking, information modeling, and searching technologies will contribute to the emergence of new scholarly practices that will enable scholars to successfully face the challenges of the information deluge era.

Below are described some innovations in scholarly practices that will be enabled, in the near future, by the cyberscholarly infrastructures described in Section 5.

11. New Discovery Practices

The availability of huge amounts of scientific information will produce a shift in the traditional scientific method: from hypothesis-driven advances to advances driven by connections and correlations found between diverse types of information resources. Discovering previously unknown and potentially useful scientific information requires the discovery of patterns within a linked scholarly record.

Given a discipline-specific linked scholarly record, a pattern is defined as a path composed of a number of (sub)datasets and article modules meaningfully connected by relations that are materialized by links.

The relations can be expressed by:

- Mathematical equations when they relate numeric fields of two (sub)datasets;
- Logical relationships among article modules and (sub)datasets;
- Semantic/rhetoric relationships among modules of articles.

A pattern describes a recurring information need in terms of relationships among some components of the scholarly record (datasets, articles) and suggests a solution. The solution consists of two or more components of the scholarly record that work together in order to satisfy a scholar's information needs.

A pattern forms a causal chain and the discovery process can take complex forms. It is expressed in high-level language and constitutes the input to a knowledge-based search engine.

In many cases, it is very useful to identify relationships among individual patterns, thus creating a connected pattern space. Such a space allows scholars to navigate from one pattern to a set of related patterns.

Given the extremely large dimensions of the modern scholarly record, a search engine can better assist scholars and scientists in finding the interesting patterns they are looking for by clearly identifying and understanding the intent behind a pattern specification. The user intent is represented in the pattern specification and contained in the query submitted to the search engine. Enabling a search engine to understand the intent of a query requires addressing the following problems: (i) precisely defining the semantics of the query intent representation; and (ii) precisely delineating the semantic boundary of the intent domain.

Two broad categories of user intent can be identified:

Targeted intent: when the desired pattern is precisely described; and

Explorative intent: when the desired pattern is described in vague terms, *i.e.*, the user does not know exactly what s/he is looking for.

12. A New Paradigm of Information Seeking: Information Exploration

In the era of scientific information deluge, the amount of information exceeds the capabilities of traditional query processing and information retrieval technologies. New paradigms of information seeking will emerge that allow scholars to:

- surf the linked scholarly information space following suitable patterns;
- explore the scholarly record searching for interesting patterns; and
- move rapidly through the linked scholarly record and identify relevant information on the move.

Information exploration is an emerging paradigm of information seeking. Exploration of a large information space is performed when scholars are searching for interesting patterns, often without knowing *a priori* what they are looking for. Exploration-based systems should help scholars to navigate the information space. In essence, a scholar supported by exploration-based systems becomes a navigator of the scientific information space.

Exploration can be conducted in two ways:

Navigational querying: in the navigational querying mode, the exploration is conducted with a specific target node in mind. In this style of exploration, the key point is the process of selecting where to go next. In order to improve the effectiveness of this process, it is important to increase the awareness of the structure of the information space.

Navigational Browsing: in the navigational browsing mode, a scholar is looking at several nodes of a linked information space in a casual way, in the hope that s/he might find something interesting. In essence, in this style of exploration the user (scholar) is not able to formulate her/his information need as a query; however, s/he is able to recognize relevant information when s/he sees it.

In this style of exploration, the most efficient strategy is a two-step approach: first, the user navigates to the topic neighborhood in a querying mode and then browses the information space within that neighborhood.

In order to increase the effectiveness of browsing, it is important to assist the user in the process of choosing between different patterns.

Browsing is distinguished from querying by the absence of a definite target in the mind of the scholar. Therefore, the distinction between browsing and querying is not determined by the actions of the scholar, or by the functionality of an information exploration system, but by the cognitive state of the scholar.

It is difficult to have a clear distinction between these two styles of exploration. Presumably, there is a continuum of user behaviors varying between knowing exactly what a user wants to find (querying) and having only an extremely vague idea of what s/he is looking for (browsing).

A wide range of exploration strategies can be defined based on the degree of target specificity in the mind of scholar. On the one extreme of the range the starting point of the exploration is the target identification and on the other extreme the starting point is the context identification.

13. Topic Maps: A Tool for Producing Conceptual Views on Top of a Linked Scholarly Record

A technology that can support scholars in finding useful information in a linked scientific information space is *Topic Maps*, a standard for connecting knowledge structures to information resources.

A Topic Map is a way of representing networked knowledge in terms of *topics*, *associations*, and *occurrences*.

- A topic is a machine-processable representation of a concept. The Topic Maps standard does not restrict the set of concepts that can be represented as topics in any way. Topics can represent any concept: in a scientific context, they can represent any scientific outcome: an article, an author, a dataset, a data mining/visualization tool, an experiment, *etc.* Typically topics are used to represent electronic resources (such as documents, web pages, and web services) and non-electronic resources (such as people or places).
- Associations represent hyper-graph relationships between topics: an article that suggests a thesis can connect to another article that supports this thesis, a data mining tool can connect to a mined dataset, *etc.*; and
- Occurrences represent information resources relevant to a particular topic.

In a Topic Map, each concept connects to another and links back to the original concept. Topics, associations, and occurrences can all be typed. Types are defined by the creator of the Topic Map(s). The definitions of allowed types constitute the ontology of the Topic Map. Each topic involved in an association is said to play a *role*, which is defined by the association *type*.

Topic Maps are a way to develop logical thinking by revealing connections and helping scholars see the lineage of an idea and how individual ideas form a larger whole.

The Topic Map can act as a high-level overview of the domain knowledge contained in a set of resources. In this way it can serve not only as a guide to locating resources for the expert, but also as a way for experts to model their knowledge in a structured way. This allows non-experts to grasp the basic concepts and their relationships before diving down into the resources that provide more detail.

Topic Maps are often described as a kind of superimposed semantic metadata layer for indexing (often dispersed and heterogeneous) information resources.

An information architecture based on Topic Maps may be said to have two layers: a knowledge layer (topic space) representing the objects in the domain being described and a content layer (resource space) holding information about these objects. With some thoughtful modeling it is even possible to create different layers of detail in a Topic Map.

Another way of looking at Topic Maps is to consider them as enablers of Knowledge Arenas, that is, virtual spaces where scholars and learners may explore what they know and what they do not know.

In fact, a Topic Map might be employed in an e-learning system to organize distributed learning resources on the web. Here the individual topics would represent digital "learning objects" like articles, video lectures, or slides.

A Topic Map can be created by a human author or automatically. The manual creation of topic maps guarantees high-quality, rich topic maps. However, even the automatic production of topic maps from a linked information space can give good results.

Topic Maps make information findable by giving every concept in the information space its own identity and providing multiple redundant navigation paths through the linked information space. These paths are semantic, and all points on the way are clearly identified with names and types that tell you what they are. This means you always know where you are. Therefore, Topic Maps can act as a GPS of the information universe.

In essence, Topic Maps can be used to create personalized semantic views, on top of a linked scholarly record that satisfies scholars' reading, learning, and research needs.

The standardization of Topic Maps is taking place under the umbrella of the ISO/IEC JTC1/SC34/WG3 Committee (ISO/IEC Joint Technical Committee 1, Subcommittee 34, Working Group 3—Document description and processing languages—Information Association). The Topic Maps (ISO/IEC 13250) reference model and data model standards are defined in a way that is independent of any specific serialization or syntax. It is desirable to have a way to arbitrarily query the data within a particular Topic Maps store. Many implementations provide a syntax by which this can be achieved (somewhat like 'SQL for Topic Maps') but the syntax tends to vary a lot between different implementations.

14. New Reading Practices

The creation of linked scientific spaces, together with the growing quantity of published articles and the limited time for reading, is increasingly modifying reading practices in two main directions: *focused* reading *vs. horizontal/explorative reading*.

14.1. Focused Reading

Due to the continuously increasing quantity of scientific articles and data and the limited time for reading, scientists strive to avoid older and less relevant literature. They want to read only the relevant parts of a small number of core articles. Therefore, they tend to narrow the literature space to be browsed (tuned vision). A number of indicators of the relevance of an article are used: indexing and citations as indicators of relevance, abstracts and literature reviews as surrogates for full papers, and social networks of colleagues as personal alerting services.

14.2. Horizontal Reading/Exploration

Another form of reading consists in surfing the linked literature space in order not to find a specific article or a core set of articles to read, but rather to find, assess, and exploit a wide range of information by scanning portions of many articles, *i.e.*, horizontal reading. Horizontal reading is the exploration of large quantities of relevant information.

14.3. Strategic Reading

Both directions lead to a new reading practice: *strategic reading*. Strategic reading is the reading of the different modules of an article in relevance order rather than narrative order.

15. New Learning Practices

A linked scientific information space has the potential for increasing the learning capacity of scholars as it supports their cognitive processes. Cognitive processes involve the creation of links between concepts. This implies the ability to create meaning by establishing patterns, relationships, and connections. A linked space enables the construction of meaningful *learning patterns* that allow the acquiring of new or modifying existing knowledge. The following of pre-constructed learning patterns makes possible the exploration and comparison of ideas, the identification or resolution of disagreements, the tracking of contributions by an individual researcher, the tracing of the lineage of an idea, *etc.*

More importantly, a linked space facilitates the establishment of meaningful connections between several information elements: interpretation, prediction, causality, consistency, prevention, (supporting/challenging) argumentation, *etc.*

16. Open Cyber-Scholarly Infrastructures

By *Cyber-scholarly Infrastructure* we mean a managed networked environment that incorporates capabilities instrumental to supporting the activities conducted by scholars. It is an enabler of an open, evolvable, and extensible learned ecosystem composed of digital libraries, publishers' repositories, institutional repositories, data repositories, data centers, and communities of scholars.

It enables interoperation between data and literature, thus creating an open, distributed, and evolvable linked scholarly record. It provides an enabling framework for data, information, and knowledge discovery, advanced literature analyses, and new scholar practices based on linking and semantic technologies. Cyber-infrastructure-enhanced discovery, analysis, reading, and learning are especially important as they encourage broadened participation and wider diversity along individual, geographical, and institutional dimensions.

In particular, future *Open Cyber-scholarship Infrastructures* should support:

- a *Scholarly Linking Environment* that:
    - provides a core set of *linking services* that create discipline-specific linked literature spaces and discipline-specific linked data spaces, connect literature spaces with data spaces, and build connections between diverse discipline-specific literature spaces.
    - supports the creation, operation, and maintenance of a core set of *linkers*. A linker is a software module that exploits encoded knowledge and metadata information about certain datasets or articles in order to build a relation between modules and/or datasets. The linking process is a two-phase process: the first phase provides assistance in locating and understanding resource capabilities; the second phase focuses on linking the identified resources.

Different types of linkers should be supported in order to implement the different types of relations between article modules and datasets. Linkers that connect modules related by a causality relationship, by a similarity relationship, by an "aboutness" relationship, or by a generic relationship; linkers that connect an article with the underlying dataset; linkers that connect a dataset with the supported articles; *etc.*

- *A Mediating Environment* that:
    - provides a core set of *intermediary services* that make the holdings of discipline-specific repositories and data centers, data archives, research digital libraries, and publisher's repositories discoverable, understandable, and (re)usable.
    - supports the creation, operation, and maintenance of *mediators*. A mediator is a software module that exploits encoded knowledge and metadata information about certain datasets or articles in order to implement an intermediary service. A core set of mediators should include: data discovery mediators, article module discovery mediators, mapping mediators, matching mediators, consistency checking mediators, data integration mediators, *etc.*
    - maintains data dictionaries, discipline-specific ontologies, and terminologies.
- *A Navigational Environment* that:
    - offers the possibility for scholars to start browsing in one dataset/article module and then navigate along links into related datasets/article modules, and/or supports search engines that crawl the linked information space by following links between datasets/article modules and provide expressive query capabilities over aggregated data.
    - maintains article module metadata registries;
    - maintains link metadata registries;
- *A Scholarly Reading and/or Learning Environment* that:
    - supports the creation, operation, and maintenance of a core set of *scholarly workflows.* Scholars should be enabled to describe an *abstract workflow* by specifying a number of abstract tasks. These tasks include identity resolution, text analysis, literature analysis, lineage analysis, reproducibility of work, repeatability of experiments, *etc.* The abstract workflow or workflow template is mapped into a *concrete workflow* using mappings that, for each task, specify a linker or a mediator, or a service to be used for its implementation. An abstract workflow is an acyclic graph in which the nodes are tasks and the edges present links that connect the output of a given task to the input of another task, specifying that the artifacts produced by the former are used by the latter.

The instantiation of a workflow results in a *scholarly reading/learning pattern*. By scholarly reading/learning pattern we mean a set of meaningfully linked article modules and datasets that support a scholarly activity (reading/learning/research). In essence, scholarly reading/learning patterns draw paths within the linked scholarly record.

- supports the creation and maintenance of reading and learning profiles in order to enable the creation of "personalized reading/learning patterns.".
- supports the creation and maintenance of *virtual information spaces (topic maps)* where scholars and learners may explore what they know and what they do not know.
- enables scholars, readers, and learners to find the scientific information they are looking for and correctly interpret it by allowing them to surf the linked scholarly record, following suitable scholarly patterns.

17. Concluding Remarks

We expect that the building of the next generation of cyber-scholarly infrastructures will have a considerable impact on:

- accelerating the transition towards an extended system of scholarly communication that allows us to create, disseminate, and sustain unprecedented new forms of scholarly inquiry by utilizing the innovative capabilities of digital technologies;
- bringing to maturity digital publishing business models that support promotion and tenure practices that systematically reward digital publishing efforts;
- making scholarly knowledge freely available to anyone and opening up the process of knowledge discovery as early as possible;
- changing the scholarly publication: making the research outcomes reproducible, replicable, and transparent; making explicit hidden aspects of knowledge production;
- overcoming the distinction between the two cultures of the contemporary scientific world (*i.e.*, the culture of data and the culture of narrative) by tightly linking datasets and narrative;
- enabling Open Scholarship;
- enabling reputation management;
- bringing into closer working alignment scholars, libraries, and publishers;
- shifting the scientific method from hypothesis-driven to data-driven discovery; and
- enabling analysis of research dynamics as well as macro-analyses of research data of interest to universities, funding bodies, academic publishers, and companies.

Future cyber-scholarly infrastructures will make Jim Gray's vision of a world in which all scientific literature and all scientific data are online and interoperating happen.

Acknowledgments: This paper has been inspired by a number of papers, listed in Appendix, that have addressed some of the issues discussed. The author is grateful to the authors of these papers.

Conflicts of Interest: The author declares no conflict of interest.

Appendix

1. Alexander, K.; Cyganiak, R.; Hausenblas, M.; Zhao, J. Describing linked datasets. In Proceedings of the Linked Data Workshop at WWW09, Madrid, Spain, 4 September 2009.
2. Altman, M.; King, G. A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine March*, April 2007.
3. Bardi, A.; Manghi, P. Enhanced publications: Data models and information systems. *LIBER Q.* **2014**, *23*, 240–273.
4. Bechhofer, S.; de Roure, D.; Gamble, M.; Goble, C.; Buchan, I. Research objects: Towards exchange and reuse of digital knowledge. In Proceedings of the Future of the Web for Collaborative Science (FWCS 2010), Raleigh, NC, USA, 6 July 2010.
5. Belhajjame, K.; Corcho, O.; Garijo, D.; Zhao, J.; Missier, P.; Palma, R.; Bechhofer, S.; García, E.; Gómez-pérez, J.M.; Klyne, G.; *et al*. Workflow-centric research objects: First class citizens in scholarly discourse. In Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012), Heraklion, Greece, 28 May 2012.
6. Bizer, C. Interlinking scientific data on a global scale. *Data Sci. J.* **2013**, *12*, GRD16–GRD112.

7. Bizer, C.; Heatth, T. Linked Data: Evolving the web into a global data space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.
8. Borgman, C. Data, disciplines, and scholarly publishing. *Learn. Publ.* **2008**, doi:10.1087/095315108X254476.
9. Bourne, P.; Clark, T.; Dale, R.; de Waard, A.; Herman, I.; Hovy, E.; Shotton, D. (Eds.) Improving the future of research communications and e-scholarship. *Dagstuhl Manif.* **2011**, doi:10.4230/DagMan.1.1.41.
10. Bourne, P. Will a biological database be different from a biological journal? *PLoS Comput. Biol.* **2005**, *1*, 179–181.
11. Shun, S.B.; Simon, J.; Li, V.U.; Sereno, B.; Mancini, C. Modeling naturalistic argumentation in research literatures: Representation and interaction design issues. *Int. J. Intell. Syst.* **2007**, doi:10.1002/int.20188.
12. Shun, S.B. Net-Centric Scholarly Discourse? Available online: http://slidesha.re/qvoqoU (accessed on 1 February 2016).
13. Clark, T.; Ciccarese, P.; Goble, C. Micropublications: A semantic model for claims, evidence, arguments and annotations in biomedical communications. *J. Biomed. Semant.* **2014**, doi:10.1186/2041-1480-5-28.
14. Burke, R.; Hammond, K.; Young, B. Knowledge-based navigation of complex information spaces. In Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, Portland, OR, USA, 4–8 August 1996.
15. Castelli, D.; Manghi, P.; Thanos, C. A vision towards scientific communication infrastructures. *Int. J. Digit. Libr.* **2013**, *13*, 3–4.
16. Decker, S. From Linked Data to Networked Knowledge. Available online: http://videolectures.net/eswc2013_decker_networked_knowledge/ (accessed on 1 February 2016).
17. De Vocht, L.; Coppens, S.; Verborgh, R.; Sande, M.V.; Mannens, E.; van de Walle, R. Discovering meaningful connections between resources in the web of data. In Proceedings of the LDOW2013, Rio de Janeiro, Brazil, 14 May 2013.
18. De Waard, A. From proteins to fairytales: Directions in semantic publishing. *IEEE Intell. Syst.* **2010**, *25*, 83–88.
19. De Waard, A.; Buckingham, S.; Carusi, A.; Park, J.; Samwald, M.S. Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. In Proceedings of the 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse, Lecture Notes in Computer Science, Washington, DC, US, 25–29 October 2009; Springer Verlag: Berlin, Germany, 2009.
20. De Waard, A.; Kircz, J. Modeling scientific research articles—Shifting perspectives and persistent issues. In Proceedings of the ELPUB 2008 Conference on Electronic Publishing, Toronto, ON, Canada, 25–27 June 2008.
21. Dillon, A.; Richardson, J.; McKnight, C. *Navigation in Hypertext: A Critical Review of the Concept*; Diaper, D., Gilmore, D., Cockton, G., Shackel, B., Eds.; Human Interaction—INTERACT'90: Amsterdam, The Netherlands, 1990; pp. 587–592.
22. Evans, J. Electronic publication and the narrowing of science and scholarship. *Science* **2008**, doi:10.1126/science.1150473.
23. Fink, L.; Fernicola, P.; Chandran, R.; Parastatidis, S.; Wade, A.; Naim, O.; Quinn, G. Word add-in for ontology recognition: Semantic enrichment of scientific literature. *Bioinfrmatics* **2010**, *11*, 103.
24. Ginsparg, P. Text in a data-centric world. In *The Fourth Paradigm: Data Intensive Scientific Discovery*; Microsoft: Redmond, WA, USA, 2009.
25. Goble, C.; de Roure, D. The Impact of Workflows on Data-centric Research. In *The Fourth Paradigm: Data Intensive Scientific Discovery*; Hey, T., Tansley, S., Tolle, K., Eds.; Microsoft Research: Redmond, WA, USA, 2009.

26. Gray, J.; Szalay, A.; Thakar, A.; Stoughton, C.; van de Berg, J. *Online Scientific Data: Curation, Publication and Archiving*; Technical Report MSR-TR- 2002-74; Microsoft Research: Redmond, WA, USA, 2002.
27. Gruber, T. Towards principles for the design of ontologies used for knowledge sharing. *Int. J. Hum. Comput. Stud.* **1995**, *43*, 907–928.
28. Halevy, A.; Franklin, M.; Maier, D. Principles of dataspace systems. In Proceedings of the PODS'06, Chicago, IL, USA, 26–28 June 2006.
29. Harmsze, F. A Modular Structure for Scientific Articles in an Electronic Environment. Ph.D. Thesis, University of Amsterdam, Amsterdam, The Netherlands, 2000.
30. Herman, I.; Clark, T.; Hovy, E.; de Waard, A. *Report on the "Future of Research Communications" Workshop*; Dragstuhl Research Online Publication Server: Dagstuhl, Germany, 15–18 August 2011; doi:10.4230/DagRep.1.8.29
31. *The Fourth Paradigm: Data Intensive Scientific Discovery*; Hey, T., Tansley, S., Tolle, K., Eds.; Microsoft Research: Redmond, WA, USA, 2009.
32. Hu, J.; Wang, G.; Lochovsky, F.; Sun, J.; Chen, Z. Understanding user's query intent with Wikipedia. In Proceedings of the WWW 2009, Madrid, Spain, 20–24 April 2009.
33. Hunter, J. Scientific models—A user—Oriented approach to the integration of scientific data and digital libraries. In Proceedings of the VALA 2006, Melbourne, Australia, 8 February 2006.
34. Idreos, S. Big data exploration. In *Big Data Computing*; Taylor and Francis: Abingdon, UK, 2013.
35. Johnsen, L. Topic maps. *J. Inf. Archit.* 1 July 2010, ISSN: 1903-7260.
36. Kavuluru, R.; Thomas, C.; Sheth, A.; Chan, V.; Wang, W.; Smith, A. An up-to-date knowledge-based literature search and exploration framework for focused bioscience domains. In Proceedings of the IHI 2012—2nd ACM SIGHIT International Health Informatics Symposium, New York, NY, USA, 28–30 January 2012.
37. Khabsa, M.; Giles, C.L. The number of scholarly documents on the web. *PLoS ONE* **2014**, *9*, e93949.
38. Kircz, J.; Harmsze, F. Modular Scenarios in the Electronic Age. In Proceedings of the Conferentie Informatiewetenschap 2000: De DoelenUtrecht, 5 April 2000; pp. 31–43.
39. Kircz, J.G. New practices for electronic publishing—New forms of the scientific paper. *Learn. Publ.* **2002**, doi:10.1087/095315102753303652.
40. Lagoze, C.; van de Sompel, H. The OAI Protocol for Object Reuse and Exchange. Available online: http://www.openarchives.org/ore (accessed on 1 February 2016).
41. Lynch, C. Jim Gray's fourth paradigm and the construction of the scientific record. In *The Fourth Paradigm: Data Intensive Scientific Discovery*; Microsoft: Redmond, WA, USA, 2009.
42. Owen, J.S.M. The Scientific Article in the Age of Digitization. Ph.D. Thesis, University of Amsterdam, Amsterdam, The Netherlands, 2005.
43. McPherson, T. Scaling vectors: Thoughts on the future of scholarly communication. *J. Electron. Publ.* **2010**, doi:10.3998/3336451.0013.208.
44. Microsoft. Patterns & Practices. Available online: https://msdn.microsoft.com/en-us/library/ff646997.aspx (accessed on 1 February 2016).
45. Microsoft. An Introduction to Topic Maps. Available online: https://msdn.microsoft.com/en-us/library/aa480048(d=printer) (accessed on 1 February 2016).
46. Nicholas, D.; Huntington, P.; Jamali, H.; Rowlands, I.; Dobrowoski, T. Viewing and Reading Behavior in a Virtual Environment. Available online: https://www.emeraldinsight.com/0001–253X.htm (accessed on 1 February 2016).
47. Nicholas, D.; Huntington, P.; Jamali, H.; Rowlands, I.; Dobrowoski, T. Characterizing and evaluating information seeking behavior in a digital environment: Spotlight on the 'Bouncer'. *Inf. Process. Manag.* **2007**, *43*, 1085–1102.
48. Paskin, N. Digital object identifier for scientific data. *Data Sci. J.* **2005**, *4*, 12–20, doi:10.2481/dsj.4.12.

49. Phelps, T.; Wilensky, R. Toward active, extensible, networked documents: Multivalent architecture and applications. In Proceedings of the ACM Digital Libraries ‘96/Bethesda, Bethesda, MD, USA, 20–23 March 1996.
50. Poggi, A.; Lembo, D.; Calvanese, D.; de Giacomo, G.; Lenzerini, M.; Rosati, R. Linking data to ontologies. *J. Data Semant. X, LNCS 4900, Pages 133–173*; Springer-Verlag: Berlin/Heidelberg, Germany, 2008.
51. Renear, A.; Sacchi, S.; Wickett, K. *Definitions of Dataset in the Scientific and Technical Literature*; ASIST: Pittsburgh, PA, USA, 2010.
52. Renear, A.; Palmer, C. Strategic reading, ontologies, and the future of scientific publishing. *Science* **2009**, *325*, 828–832.
53. Seringhaus, T.; Gerstein, M. Publishing perishing? Towards tomorrow’s information. *BMC Bioinform.* **2007**, *8*, 17.
54. Shotton, D. Semantic publishing: The coming revolution in scientific Journal publishing. *Learn. Publ.* **2009**, doi:10.1087/2009202.
55. Simon, B.; Miklos, Z.; Nejdl, W.; Sintek, M.; Salvachua, J. Smart Space for Learning: A Mediation Infrastructure for Learning Services. Available online: https://wwwconference.org/www2003/cdrom/papers/alternate (accessed on 1 February 2016).
56. Taylor, I.; Gannon, D.; Shields, M. (Eds.) *Workflows for E-Science*; Springer –Verlag: London, UK, 2009.
57. Tenopir, C.; King, D.; Edwards, S.; Wu L. Electronic journals and changes in scholarly article seeking and reading patterns. *Aslib Proc.* **2009**, *61*, 5–32, doi:10.1108/00012530910932267.
58. Thearling, K. An Introduction to Data Mining. Available online: http://www.thearling.com/dmintro/dmintro_2.htm (accessed on 1 February 2016).
59. Vieu, L. On the Semantics of Discourse Relations. Available online: http://www.irit.fr/publis/LILAC/V-DRSemantics-CID11.pdf (accessed on 1 February 2016).
60. Waterworth, J.; Chignell, M. A Model for Information Exploration. Available online: http://www8.informatik.umu.se/~jwworth/infomodel.pdf (accessed on 1 February 2016).
61. White, C. Data exploration and discovery: A new approach to analytics. *BI Res.* 2013, doi:not available.
62. Woutersen-Windhouwer, S.; Brandsma, R.; Verhaar, P.; Hogenaar, A.; Hoogerwerf, M.; Doorenbosch, P.; Durr, E.; Ludwig, J.; Schmidt, B.; Sierman, B. *Enhanced Publications*; Vernooy-Gerritsen, M., SURF Foundation, Eds.; Amsterdam University Press: Amsterdam, The Netherlands, 2009.

References

1. Thanos, C. The future of digital scholarship. *Proced. Comput. Sci.* **2014**, *38*, 22–27. [CrossRef]
2. Porter, B.; Souther, A. *Knowledge-Based Information Retrieval*; AAAI Technical Report FS-99-02; Portland, OR, US, 1999.
3. Franklin, M.; Halevy, A.; Maier, D. From databases to dataspaces: A new abstraction for information management. In *SIGMOD Record*; ACM: New York, NY, USA, 2005; Volume 34, pp. 27–33. [CrossRef]
4. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data—The story so far. *Int. J. Semant. Web Inf. Syst.* **2009**, *5*, 1–22. [CrossRef]