Digital Libraries: From Digital Resources to Challenges in Scientific Data Sharing and Re-use

Maristella Agosti, Nicola Ferro and Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
{maristella.agosti, nicola.ferro, gianmaria.silvello}@unipd.it

1 Introduction

Digital libraries and digital archives are the information management systems for storing, indexing, searching, accessing, curating and preserving digital resources which manage our cultural and scientific knowledge heritage (KH). They act as the main conduits for widespread access and exploitation of KH related digital resources by engaging many different types of users, ranging from generic and leisure to students and professionals.

In this chapter, we describe the evolution of digital libraries and archives over the years, starting from *Online Public Access Catalog (OPAC)*, passing through monolithic and domain specific systems, up to service-oriented and component-based architectures. In particular, we present some specific achievements in the field: the DELOS Reference Model and the DelosDLMS, which provide a conceptual reference and a reference implementation for digital libraries; the FAST annotation service, which defines a formal model for representing and searching annotations over digital resources as well as a RESTful Web service implementation of it; the NESTOR model for digital archives, which introduces an alternative model for representing and managing archival resources in order to enhance interoperability among archives and make access to them faster; and, the CULTURA environment, which favours user engagement over multimedia digital resources.

Finally, we discuss how digital libraries and archives are a key technology for facing upcoming challenges in data sharing and re-use. Indeed, due to the rapid evolution of the nature of research and scientific publishing which are increasingly data-driven, digital libraries and archives are also progressively addressing the issues of managing scientific data. In this respect, we focus on some key building blocks of this new vision: data citation to foster accessibility to scientific data as well as transparency and verifiability of scientific claims, reproducibility in science as an exemplar showcase of how all these methods are indispensable for addressing fundamental challenges, and keyword-based search over relation/structured data to empower natural language access to scientific data.

2 Evolution of Digital Libraries

The term "digital libraries" corresponds to a very complex notion with several diverse aspects and it cannot be captured by a simple definition. Indeed, the term is used to refer to systems that are very heterogeneous in scope and provide very different functionalities [28]. These systems span from digital object and metadata repositories, reference-linking systems, archives, and content administration systems, to complex systems that integrate advanced services. Furthermore, digital libraries represent the meeting point of many disciplines and research fields – i.e. database management, information retrieval, library and information sciences, document and information systems, the Web, information visualization, artificial intelligence, human-computer interaction, and others [49].

Initially, digital libraries were almost monolithic systems, each one built for a specific kind of information resources – e.g. images or videos – and with very specialized functions developed ad-hoc for those contents. This approach caused a flourishing of systems where the very same functions were developed and redeveloped many times from scratch. Moreover, these systems were confined to the realm of traditional libraries, since they were the digital counterpart of the latter, and they had a kind of static view of their role, which was document-centric rather than user-centric.

In the 1980s the most advanced library automation systems were designed to include procedures also able to collect log data that were used to manage the system itself, and especially to monitor the usage of system search facilities by users, where the search facility which was designed for user search and access to catalog data was an OPAC [48]; some OPACs were reachable in a distributed environment: an example of such a system is the DUO OPAC system from the early 1990s [15]. Towards the end of the 1980s/beginning of 1990s it became apparent that a library automation system could not only manage catalog data or metadata describing physical objects, but also digital files representing physical objects. Digital libraries started to be seen as increasingly user-centered systems, where the original content management task is partnered with new communication and cooperation tasks.

In this evolving scenario, the design and development of effective services which foster cooperation among users and the integration of heterogeneous information resources becomes a key factor. Digital libraries are thus no longer perceived as isolated systems but, on the contrary, as systems that need to cooperate with each other to improve the user experience and give personalized services. Nowadays, there are several accepted conceptions of digital libraries:

- *User-centric systems*: Digital libraries as user-centered information infrastructures able to support content management tasks together with tasks devoted to communication and cooperation. Although they are still places where information resources can be stored and made available to end users, recent design and development efforts move in the direction of transforming them into infrastructures able to support the user in different information centric activities.

– *Dynamic interactions*: Digital libraries as dynamic forms of facilitation of communication, collaboration and other forms of interaction among scientists, researchers and the general public.
– *Large capabilities*: Digital libraries as systems able to handle distributed multimedia document collections, sensor data, mobile information, and pervasive computing services.

Digital libraries have contributed to supporting the creation of innovative applications and services to access, share and search our cultural KH. In this context, another key feature we have to consider to understand the world of digital libraries is that they have to take into account several distributed and heterogeneous information sources with different community background and different information objects ranging from full content of digital objects to the metadata describing them. These objects can be exchanged between distributed systems or they can be aggregated and accessed by users with distinct information needs and living in different countries. Indeed, one of the most important contributions of digital libraries is to make available collections of digital resources from different cultural institutions such as *libraries*, *archives* and *museums*, to make them accessible in different languages and to provide advanced services over them. We have to consider that the above mentioned institutions are different from several point-of-views: their internal organization has different peculiarities, the resources they collect and manage have different structure and nature, these resources are described with different means and for different purposes, their users have different information needs and require different methods to access the resources. Thus, digital libraries are heterogeneous systems with peculiarities and functions that range from data representation to data exchange and data management. Furthermore, digital libraries are meaningful parts of a global information network which includes scientific repositories, curated databases and commercial providers. All these aspects need to be taken into account and balanced to support final users with effective and interoperable information systems.

A fundamental role of digital libraries therefore is to provide data models, protocols, applications and services to handle all these resources, all the while preserving their characteristics and addressing the issues related to their differences.

3 Models and Services for Digital Libraries

Digital libraries have shaped the way for accessing our cultural heritage and have become primary knowledge conduits thanks to the development of formal and conceptual models of what digital libraries are, such as the DELOS Reference Model [28] and the *Streams, Structures, Spaces, Scenarios, Societies (5S)* model [45]. These models have then been specialised to specific domains, such as archives [41], and to specific services, such as digital annotations [10]. Finally, thanks to the recent development of semantic technologies, it has been

also possible to provide formal mappings between the DELOS Reference Model and the 5S models to improve interoperability among digital libraries [13]. In the following, we briefly present our main contributions in this context.

3.1 DELOS Reference Model and DelosDLMS

The DELOS Reference Model approaches the problem of modelling the digital library universe by highlighting six domains or main concepts [28], which are at the core of what digital libraries are and what their purpose is:

- *Content*: the data and information that digital libraries handle and make available to their users;
- *User*: the actors (whether human or not) entitled to interact with digital libraries;
- *Functionality*: the services that digital libraries offer to their users;
- *Quality*: the parameters that can be used to characterize and evaluate the content and behaviour of digital libraries;
- *Policy*: a set of rules that govern the interaction between users and digital libraries;
- *Architecture*: a mapping of the functionality and content offered by a digital library onto hardware and software components.

These six domains represent the high level containers that help organize the DELOS Reference Model. For each of these concepts, the fundamental entities and their relationships are clearly defined and discussed. Note that these six domains are not separate, but, on the contrary, are strongly inter-related; the entities within one domain are often related to or influenced by the entities in other domains.

Moreover, the DELOS Reference Model distinguishes between three different "systems" which constitute the digital library universe and rely on the six domains introduced above for their definition:

- *Digital Library (DL)*: an organisation, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialised functionality on that content, of measurable quality and according to codified policies.
- *Digital Library System (DLS)*: a software system that is based on a defined (possibly distributed) architecture and provides all functionality required by a particular Digital Library. Users interact with a Digital Library through the corresponding Digital Library System.
- *Digital Library Management System (DLMS)*: a generic software system that provides the appropriate software infrastructure both (i) to produce and administer a Digital Library System incorporating the suite of functionality considered fundamental for Digital Libraries and (ii) to integrate additional software offering more refined, specialised or advanced functionality.

The three systems are at different levels of abstraction and constitute a kind of hierarchy: at the more general level there is the notion of DL, which is what is actually perceived by the end-users and what they interact with; in-between, there is the DLS, which mainly concerns system designers and administrators who have to instantiate and manage it; at the lower level, there is the DLMS, which typically interests system developers who implement the actual components that are used by the upper layers.

3.2 The FAST Annotation Model

The *Flexible Annotation Service Tool (FAST)* [33] covers many of the uses and applications of annotations, since it is able to represent and manage annotations which range from metadata to full content; its flexible and modular architecture makes it suitable for annotating general Web resources as well as digital objects managed by different digital library systems; the annotations themselves can be complex multimedia compound objects, with a varying degree of visibility which ranges from private to shared and public annotations and different access rights. The FAST annotation service has proven its flexibility and adaptability to different applicative contexts in many different ways. It has been integrated into the DelosDLMS [3], the prototype of the next generation digital library system developed by DELOS, and in the CULTURA environment [6,5].

FAST adopts and implements the formal model for annotations proposed in [10]. Annotations are compound multimedia objects constituted by different *signs of annotation* that materialize the annotation itself. For example, we can have *textual signs*, which contain the textual content of the annotation, *image signs*, if the annotation is made up of images, and so on. In turn, each sign is characterized by one or more *meanings of annotation* that specify the semantics of the sign. Moreover, an annotation is uniquely identified by a handle, which usually takes the form of a pair (namespace, identifier), where the namespace provides logical grouping of the identifiers, it has a scope which defines its visibility, and it can be shared with different groups of users.

Annotations can be linked to digital objects with two main types of links: (1) *annotate link* an annotation annotates a digital object, which can be either a document or another annotation; (2) *relate-to link* an annotation relates to a digital object, which can be either a document or another annotation. The hypertext between annotations and annotated objects can be exploited for providing alternative navigation and browsing capabilities. In addition, it can span and cross the boundaries of the single digital library and also related to Web resources. Most importantly, this hypertext can be exploited to develop advanced search functionalities [9]. Based on the proposed formal model, we developed a fully-fledged search model, mixing exact match and best match queries, paired with an intuitive query language expressed in the *Contextual Query Language (CQL)* syntax [32]. In this way, we can not only search for annotations, by means of a mix of full text and queries based on the structure, but we can also retrieve annotated resources thanks to the hypertext that allows us to pass from the found annotations to the annotated resources.

3.3 NESTOR: A Model for Digital Archives

Digital archives are one of the pillars of our cultural heritage and, thanks to technologies such as the digital libraries, they are increasingly opening up to end-users by focusing on usability, accessibility, and findability of the resources they manage.

Archives represent the trace of the activities of a physical or juridical person in the course of their business which is preserved because of their continued value over time. Archives and archival descriptions (i.e. metadata) are modeled by using a hierarchical structure, which expresses the relationships and dependency links between the records of the archive.

In recent years, archival descriptions have moved on-line and there have been increasing calls for reconsidering their presentation based on user studies. Indeed, from this new point-of-view, the XML standard for digital description of archives such as the *Encoded Archival Description (EAD)* seriously constrains user orientation of archives; but with EAD several important digital archives operations are not possible: (i) the user cannot access a specific item on-the-fly, instead we have to define fixed access points to the archival hierarchy; (ii) the user cannot reconstruct the context of an item without browsing the archival hierarchy; (iii) we cannot present the users with selected items from an archive, instead we have to give them the archive as a whole.

To tackle these issues, we proposed to model archives through the *NEsted SeTs for Object hieRarchies (NESTOR)* [39,41] – i.e. a set-based data model allowing for the representation of hierarchical relationships between objects through the inclusion property between sets – which opened-up new ways of representing and handling archival resources. The NESTOR model is defined by two set-based data models: The *Nested Set Model (NS-M)* and the *Inverse Set Data Model (INS-M)*. These models are defined in the context of set theory as a collection of subsets, their properties have been formally proved as well as their equivalence to the tree in terms of expressive power [37,41,12].

The most intuitive way to understand how these models work is to relate them to the tree. In the NS-M each node of the tree is mapped into a set, where child nodes become *proper subsets* of the set created from the parent node. Every set is subset of at least one set; the set corresponding to the tree root is the only set without any superset and every set in the hierarchy is subset of the root set. The external nodes are sets with no subsets. The tree structure is maintained thanks to the nested organization and the relationships between the sets are expressed by the set inclusion order.

The second data model is the INS-M where each node of the tree is mapped into a set, where each parent node becomes a subset of the sets created from its children. The set created from the tree's root is the only set with no subsets and the root set is a proper subset of all the sets in the hierarchy. The leaves are the sets with no supersets and they are sets containing all the sets created from the nodes composing tree path from a leaf to the root.

NESTOR is particularly well-suited for advancing user-orientation of archives because it allows for exposing archival data as Linked Data on the Web [40], thus

augmenting the understandability of these data. Furthermore, NESTOR can be adopted for "socializing the archives" [41] by means of annotations [38] such that available resources can be augmented with user-generated content which then provides alternative access points for searching and browsing resources. Furthermore, NESTOR has been realized by means of three alternative in-memory dictionary-based data structures, which have been proved to be highly competitive with state-of-the-art solutions for accessing XML data by considering pre-processing and query execution time and memory occupation [52,43,42].

3.4 User Engagement: The CULTURA Environment

The main aim of the CULTURA environment was to create a *Virtual Research Environment (VRE)* in which users with a range of different backgrounds and expertise can collaboratively explore, interrogate, interact with, and interpret complex and diverse digital cultural heritage collections [6]. The CULTURA environment is a VRE that pushed forward the frontiers of technology in the creation of community and content aware interfaces to digital humanities collections.

The CULTURA environment adopts a service-oriented approach to offer a rich and engaging experience for different user categories, which range from academic and professional users to the general public. The services are conceived and developed to be applicable to a wide variety of document collections [14]. The potential generality of the environment is demonstrated by the fact that the environment supports different use cases; one of those is represented by the IPSA collection, a digital archive of illuminated manuscripts, while the other major archive is the 1641 Depositions, which is a collection of noisy text documents, mainly of a legal nature, dating from the 17th Century.

In both collections, the managed digital objects – "either scanned illuminated manuscripts or legal documents" – are described by appropriate metadata, according to a traditional record-centric approach. The goal of the environment was to exploit an improved user engagement and interaction with the managed artifacts in order to semantically enrich them with a superimposed layer of user-provided information. This required a move from a traditional record-centric approach to a resource-centric one, opened towards *Linked Open Data (LOD)* and a better sharing of resources.

The history of art provides a fertile ground for research into semantically enriched metadata and LOD; indeed, in history of art the main way to produce new knowledge is to reveal connections between different items (illuminations, pictures, frescos) that can cast new light on an artist, an artistic movement or an art-historical period. The most valuable connections are the unexpected ones linking elements that may seem to have very few features in common [54]. Therefore, it was decided that the central tool for allowing researchers in history of art to discover new knowledge and unveil new links and relationships among resources would be the semantic annotation tool, called FAST-CAT. This software enables semantically-typed links to be superimposed over the managed digital objects, the traditional record-centric metadata, and Web resources in general.

Besides being semantically typed, these links can include fully-fledged multimedia content, which allows for rich description and explanation of the link and provides added value to both specialist users and the general public.

Both the FAST annotation model [10] and the CAT model and tool (FAST-CAT) [33] have been applied to the CULTURA environment and they provided adaptive and personalized access to the IPSA historical collection. FAST-CAT has been integrated into the environment in order to provide users with an additional means of interacting with the portal, as well as for providing feedback on CULTURA user model that stores user interests. It is the belief of the authors that FAST-CAT has huge potential as an annotation tool within the digital humanities field. Indeed, it demonstrates the feasibility of transitioning from a traditional digital archive with a record-centric approach, towards a resource-centric one with semantically enriched information provided by actively engaging users via digital annotations [59]. The process of discovering new unexpected connections among cultural heritage artifacts, i.e. a process that can be defined as serendipity, enabled by the FAST annotation model, is especially encouraged by the LOD paradigm where meaningful links between entities allow us to move across diverse and apparently unrelated knowledge domains.

4 Data-Driven Digital Libraries

The role of data is going to become central in DL as well as in the other fields of computer science. In this section we explore the role of DL in the context of reproducibility in science and the relationships between DL and data citation.

4.1 Reproducibility

Computer science is particularly active in reproducibility, as witnessed by the recent *Association for Computing Machinery (ACM)* policy on result and artifact review and badging[1]. For instance, the database community started an effort called "SIGMOD reproducibility" [44] "to assist in building a culture of sharing results, code, and scripts of database research"[2]. Since 2015, the *European Conference in IR (ECIR)* [46,35], allocated a whole paper track on reproducibility and in 2015 the RIGOR workshop at SIGIR was dedicated to this topic [16]. Moreover, in 2016 the "Reproducibility of Data-Oriented Experiments in e-Science" seminar was held in Dagstuhl (Germany) [1] bringing together researchers from different fields of computer science with the goal "to come to a common ground across disciplines, leverage best-of-breed approaches, and provide a unifying vision on reproducibility" [36,34].

In recent years, the nature of research and scientific publishing has been rapidly evolving and progressively relying on data to sustain claims and provide experimental evidence for scientific breakthroughs [47]. The preservation, management, access, discovery and retrieval of research data are topics of utmost

[1] `https://www.acm.org/publications/policies/artifact-review-badging`
[2] `http://db-reproducibility.seas.harvard.edu/`

importance as witnessed by the great deal of attention they are receiving from the scientific and publishing communities [24]. Along with the pervasiveness and availability of research data, we are witnessing the growing importance of citing these data. Indeed, data citation is required to make results of research fully available to others, provide suitable means to connect publications with the data they rely upon [61], give credit to data creators, curators and publishers [23], and enabling others to better build on previous results and to ask new questions about data [22].

Even though *Information Retrieval (IR)* has a long tradition in ensuring that the due scientific rigor is guaranteed in producing experimental data, it does not have a similar tradition in managing and taking care of such valuable data [8,31]. This represents a serious obstacle for tackling the above mentioned challenges. For example, there is a lack of commonly agreed formats for modeling and describing the experimental data as well as almost no metadata (descriptive, administrative, copyright, etc.) for annotating and enriching them. The semantics of the data themselves is often not explicit and it is demanded to the scripts typically used for processing them, which are often not well documented, rely on rigid assumptions on the data format or even on side effects in processing the data. Finally, IR lacks a commonly agreed mechanism for citing and linking data to the papers describing them [57].

There have been early examples of systems to manage IR experimental data, such as EvaluatIR [17] and *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*[3] [11,7], but they have not been designed with reproducibility and/or data citation as goals. More recently, steps toward more fine grained models and systems have been proposed, as for example LOD-DIRECT[4] [56] which uses semantic Web and LOD technologies to model IR evaluation data and make them linkable, or nanopublications for IR evaluation [50].

The situation is even more severe in the context of keyword-based search over relational databases, which is a key technology for lowering the barriers of access to the huge amounts of data managed by databases. It is an extremely difficult and open challenge [2] since it comes up against the "conflict of impedance" between vague and imprecise user information needs and rigorously structured data, allowing users to express their queries in natural language against a potentially unknown database.

Even if there have been attempts to reproduce state-of-the-art solutions and provide shared benchmarks [29], we still need to move beyond the evaluation of keyword search components in isolation or not related to the actual user needs, and, instead, to consider the whole system, its constituents, and their interrelations with the ultimate goal of supporting actual user search tasks [21,20]. Moreover, there is a lack of commonly shared open source platforms implementing state-of-the-art algorithms for keyword-based access to relational data as,

[3] `http://direct.dei.unipd.it/`
[4] `http://lod-direct.dei.unipd.it/`

for example, Terrier[5] in the information retrieval field, and we just started to move towards providing open source implementations of these algorithms[6] [18].

4.2 Data Citation

The practice of citation is foundational for scientific advancement and the propagation of knowledge and it is one of the basic means on which scholarship and scientific publishing rely. In recent years, the nature of research and scientific publishing has been rapidly evolving and progressively relying on data to sustain claims and provide experimental evidence for scientific breakthroughs [47]. The preservation, management, access, discovery and retrieval of research data are topics of utmost importance as witnessed by the great deal of attention they are receiving from the scientific and publishing communities [24,25].

Along with the pervasiveness and availability of research data, we are witnessing the growing importance of citing these data. Indeed, data citation is required to make results of research fully available to others, provide suitable means to connect publications with the data they rely upon, give credit to data creators, curators and publishers, and enable others to better build on previous results and to ask new questions about data [58]. Furthermore, data citation plays a central role for providing better transparency and reproducibility in science, a challenge taken up by several fields.

In the traditional context of printed material, the practice of citation has been evolving and adapting over the centuries [24] reaching a stable and reliable state; nevertheless, traditional citation methods and practices cannot be easily applied for citing data. Indeed, citing data poses new and significant challenges, such as the use of heterogeneous data models and formats requiring different methods to manage, retrieve and access the data; the transience of data calling for versioning and archiving methods and systems; the necessity to cite data at different levels of coarseness requiring methods to identify, select and reference specific subsets of data; and the necessity to automatically generate citations to data because we cannot assume that the people citing the data understand the complexity of a dataset, know how data should be cited in a specific context, or select relevant information to form a complete and correct citation.

As described in [26], from the computational perspective the problem of data citation can be formulated as follows: "Given a dataset D and a query Q, generate an appropriate citation C". Several of the existing approaches to address this problem allow us to reference datasets as a single unit having textual data serving as metadata source, but as pointed out by [51] most data citations "can often not be generated automatically and they are often not machine interpretable". Until now, the problem of how to cite a dataset at different levels of coarseness, to automatically generate citations and to create human- and machine-readable citations has been tackled only by a few working systems [51,27,53,30,55]. From

[5] `http://www.terrier.org/`
[6] `https://bitbucket.org/ks-bd-2015-2016/ks-unipd`

these experiences, it clearly emerges that data citation is a compound and complex problem and a "one size fits all" system to address it does not exist, yet. As a consequence, within the context of data citation, there are several open issues and research directions we can take into account:

- *Citation identity and containment*: This problem refers to the necessity of uniquely identifying a citation to data and of being able to discriminate between two citations referring to different data or different versions of the same data and between two different citations referring to the same data.
- *Versioning*: One of the main differences between traditional citations and data citations is that data may not be fixed, but may evolve through time; indeed, new data may be added to a dataset, some changes may occur, some mistakes may be fixed or new information may be added. A citation needs to ensure that the data a citation uses is identical to that cited.
- *Provenance*: Provenance information plays a central role because we may need to reconstruct the chain of ownership of a data object or the chain of modifications that occurred to it in order to produce a reliable citation [30]. New solutions have to be provided to integrate data citation with currently employed systems controlling and managing the data workflow.
- *Supporting scientific claims*: Scientific claims are often based on evidence gathered from data. They could be related to a single datum or to multiple data coming from the same source or from different sources. Data citation can be used to support such claims and to provide a means to verify their reliability.

References

1. Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. In: Freire, J., Fuhr, N., Rauber, A. (eds.) Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. Dagstuhl Reports, Volume 6, Number 1, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Germany (2016)
2. Abadi, D., Agrawal, R., Ailamaki, A., Balazinska, M., Bernstein, P.A., Carey, M.J., Chaudhuri, S., Dean, J., Doan, A., Franklin, M.J., Gehrke, J., Haas, L.M., Halevy, A.Y., Hellerstein, J.M., Ioannidis, Y.E., Jagadish, H.V., Kossmann, D., Madden, S., Mehrotra, S., Milo, T., Naughton, J.F., Ramakrishnan, R., Markl, V., Olston, C., Ooi, B.C., Rè, C., Suciu, D., Stonebraker, M., Walter, T., Widom, J.: The Beckman Report on Database Research. ACM SIGMOD Record 43(3), 61–70 (September 2014)
3. Agosti, M., Berretti, S., Brettlecker, G., del Bimbo, A., Ferro, N., Fuhr, N., Keim, D., Klas, C.P., Lidy, T., Milano, D., Norrie, M., Ranaldi, P., Rauber, A., Schek, H.J., Schreck, T., Schuldt, H., Signer, B., Springmann, M.: DelosDLMS – the Integrated DELOS Digital Library Management System. In: Thanos et al. [60], pp. 36–45
4. Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.): Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009). Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany (2009)

5. Agosti, M., Conlan, O., Ferro, N., Hampson, C., Munnelly, G.: Interacting with Digital Cultural Heritage Collections via Annotations: The CULTURA Approach. In: Marinai, S., Marriot, K. (eds.) Proc. 13th ACM Symposium on Document Engineering (DocEng 2013). pp. 13–22. ACM Press, New York, USA (2013)
6. Agosti, M., Conlan, O., Ferro, N., Hampson, C., Munnelly, G., Ponchia, C., Silvello, G.: Enriching Digital Cultural Heritage Collections via Annotations: The CULTURA approach. In: Greco, S., Picariello, A. (eds.) 22nd Italian Symposium on Advanced Database Systems, SEBD 2014. pp. 319–326 (2014)
7. Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., Silvello, G.: DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012). Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany (2012)
8. Agosti, M., Di Nunzio, G.M., Ferro, N.: The Importance of Scientific Data Curation for Evaluation Campaigns. In: Thanos et al. [60], pp. 157–166
9. Agosti, M., Ferro, N.: Annotations as Context for Searching Documents. In: Crestani, F., Ruthven, I. (eds.) Proc. 5th International Conference on Conceptions of Library and Information Science – Context: nature, impact and role (CoLIS 5). pp. 155–170. Lecture Notes in Computer Science (LNCS) 3507, Springer, Heidelberg, Germany (2005)
10. Agosti, M., Ferro, N.: A Formal Model of Annotations of Digital Content. ACM Transactions on Information Systems (TOIS) 26(1), 3:1–3:57 (2008)
11. Agosti, M., Ferro, N.: Towards an Evaluation Infrastructure for DL Performance Evaluation. In: Tsakonas, G., Papatheodorou, C. (eds.) Evaluation of Digital Libraries: An insight into useful applications and methods. pp. 93–120. Chandos Publishing, Oxford, UK (2009)
12. Agosti, M., Ferro, N., Silvello, G.: The NESTOR Framework: Manage, Access and Exchange Hierarchical Data Structures. In: Martoglia, R., Bergamaschi, S., Lodi, S., Sartori, C. (eds.) Proc. 18th Italian Symposium on Advanced Database Systems (SEBD 2010). pp. 242–253. Società Editrice Esculapio, Bologna, Italy (2010)
13. Agosti, M., Ferro, N., Silvello, G.: Digital Library Interoperability at High Level of Abstraction. Future Generation Computer Systems (FGCS) 55, 129–146 (2016)
14. Agosti, M., Manfioletti, M., Orio, N., Ponchia, C.: Evaluating the Deployment of a Collection of Images in the CULTURA Environment. In: Proc. of the International Conference on Theory and Practice of Digital Libraries, TPDL 2013. Lecture Notes in Computer Science, vol. 8092, pp. 180–191. Springer (2013)
15. Agosti, M., Masotti, M.: Design of an OPAC Database to Permit Different Subject Searching Accesses in a Multi-disciplines Universities Library Catalogue Database. In: Belkin et al. [19], pp. 245–255
16. Arguello, J., Crane, M., Diaz, F., Lin, J., Trotman, A.: Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). SIGIR Forum 49(2), 107–116 (December 2015)
17. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: EvaluatIR: an Online Tool for Evaluating and Comparing IR Systems. In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (eds.) Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009). p. 833. ACM Press, New York, USA (2009)
18. Badan, A., Benvegnù, L., Biasetton, M., Bonato, G., Brighente, A., Cenzato, A., Ceron, P., Cogato, G., Marchesin, S., Minetto, A., Pellegrina, L., Purpura, A.,

Simionato, R., Soleti, N., Tessarotto, M., Tonon, A., Vendramin, F., Ferro, N.: Towards open-source shared implementations of keyword-based access systems to relational data. In: Ferro, N., Guerra, F., Ives, Z., Silvello, G., Theobald, M. (eds.) Proc. 1st EDBT/ICDT Workshop on Keyword-based Access and Ranking at Scale (KARS 2017). CEUR Workshop Proc. (CEUR-WS.org), ISSN 1613-0073 (2017)
19. Belkin, N.J., Ingwersen, P., Mark Pejtersen, A., Fox, E.A. (eds.): Proc. 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1992). ACM Press, New York, USA (1992)
20. Bergamaschi, S., Ferro, N., Guerra, F., Silvello, G.: A Perspective Look at Keyword-based Search Over Relation Data and its Evaluation. In: Atzeni, P., Lenzerini, M., Lembo, D., Torlone, R. (eds.) Proc. 23rd Italian Symposium on Advanced Database Systems (SEBD 2015) (2015)
21. Bergamaschi, S., Ferro, N., Guerra, F., Silvello, G.: Keyword-based Search over Databases: A Roadmap for a Reference Architecture Paired with an Evaluation Framework. LNCS Transactions on Computational Collective Intelligence (TCCI) 9630, 1–20 (2016)
22. Borgman, C.L.: The Conundrum of Sharing Research Data. JASIST 63(6), 1059–1078 (2012), http://dx.doi.org/10.1002/asi.22634
23. Borgman, C.L.: Why are the Attribution and Citation of Scientific Data Important? In: Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop. National Academy of Sciences' Board on Research Data and Information. pp. 1–8. National Academies Press: Washington DC (2012)
24. Borgman, C.L.: Big Data, Little Data, No Data. MIT Press (2015)
25. Brase, J., Socha, Y., Callaghan, S., Borgman, C.L., Uhlir, P.F., Carroll, B.: Research Data Management: Practical Strategies for Information Professionals, chap. Data Citation: Principles and Practice, pp. 167–186. Purdue University Press (2014)
26. Buneman, P., Davidson, S.B., Frew, J.: Why data citation is a computational problem. Communications of the ACM (CACM) 59(9), 50–57 (2016)
27. Buneman, P., Silvello, G.: A Rule-Based Citation System for Structured and Evolving Datasets. IEEE Data Eng. Bull. 33(3), 33–41 (2010), http://sites.computer.org/debull/A10sept/buneman.pdf
28. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model. Foundations for Digital Libraries. ISTI-CNR at Gruppo ALI, Pisa, Italy, http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf (December 2007)
29. Coffman, J., Weaver, A.C.: An Empirical Performance Evaluation of Relational Keyword Search Techniques. IEEE Transactions on Knowledge and Data Engineering (TKDE) 1(26), 30–42 (2014)
30. Davidson, S.B., Deutsch, D., Tova, M., Silvello, G.: A Model for Fine-Grained Data Citation. In: 8th Biennial Conference on Innovative Data Systems Research (CIDR 2017) (2017)
31. Dussin, M., Ferro, N.: Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In: Agosti et al. [4], pp. 63–74
32. Ferro, N.: Annotation Search: The FAST Way. In: Agosti et al. [4], pp. 15–26
33. Ferro, N.: The FAST Annotation Service. In: De Antonellis, V., Castano, S., Catania, B., Guerrini, G. (eds.) Proc. 17th Italian Symposium on Advanced Database Systems (SEBD 2009). pp. 169–176. Seneca Edizioni, Torino, Italia (2009)

34. Ferro, N.: Reproducibility Challenges in Information Retrieval Evaluation. ACM Journal of Data and Information Quality (JDIQ) 8(2), 8:1–8:4 (January 2017)
35. Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.): Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016). Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany (2016)
36. Ferro, N., Fuhr, N., Järvelin, K., Kando, N., Lippold, M., Zobel, J.: Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". SIGIR Forum 50(1) (June 2016)
37. Ferro, N., Silvello, G.: The NESTOR Framework: How to Handle Hierarchical Data Structures. In: Agosti et al. [4], pp. 215–226
38. Ferro, N., Silvello, G.: FAST and NESTOR: How to Exploit Annotation Hierarchies. In: Agosti, M., Esposito, F., Thanos, C. (eds.) Digital Libraries. Proceedings of the Sixth Italian Research Conference (IRCDL 2010). pp. 55–66. Springer-Verlag, Heidelberg, Germany (2010)
39. Ferro, N., Silvello, G.: The NESTOR Model: Properties and Applications in the Context of Digital Archives. In: Mecca, G., Greco, S. (eds.) Proc. 19th Italian Symposium on Advanced Database Systems (SEBD 2011). pp. 274–285. Università della Basilicata, Italy (2011)
40. Ferro, N., Silvello, G.: Modeling Archives by means of OAI-ORE. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) Digital Libraries and Archives - Proc. 8th Italian Research Conference (IRCDL 2012). Communications in Computer and Information Science (CCIS) 354, Springer, Heidelberg, Germany (2013)
41. Ferro, N., Silvello, G.: NESTOR: A Formal Model for Digital Archives. Information Processing & Management 49(6), 1206–1240 (2013)
42. Ferro, N., Silvello, G.: Descendants, Ancestors, Children and Parent: A Set-Based Approach to Efficiently Address XPath Primitives. Information Processing & Management 52(3), 399–429 (2016)
43. Ferro, N., Silvello, G.: Fast Access to XML Data: A Set-based Approach. In: Paolini, P., Bochicchio, M.A., Mecca, G. (eds.) Proc. 24th Italian Symposium on Advanced Database Systems (SEBD 2016) (2016)
44. Freire, J., Bonnet, P., Shasha, D.: Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012. pp. 593–596 (2012), http://doi.acm.org/10.1145/2213836.2213908
45. Gonçalves, M.A., Fox, E.A., Watson, L.T., Kipp, N.A.: Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Transactions on Information Systems (TOIS) 22(2), 270–312 (April 2004)
46. Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.): Advances in Information Retrieval. Proc. 37th European Conference on IR Research (ECIR 2015). Lecture Notes in Computer Science (LNCS) 9022, Springer, Heidelberg, Germany (2015)
47. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, USA (2009)
48. Hildreth, C.R. (ed.): The online catalogue: developments and directions. Library Association, London, UK (1989)
49. Ioannidis, Y.E.: Digital Libraries at a Crossroads. International Journal on Digital Libraries 5(4), 255–265 (2005)
50. Lipani, A., Piroi, F., Andersson, L., Hanbury, A.: An Information Retrieval Ontology for Information Retrieval Nanopublications. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation – Multilinguality, Multimodality, and Interaction. Proceedings of

the Fifth International Conference of the CLEF Initiative (CLEF 2014). pp. 44–49. Lecture Notes in Computer Science (LNCS) 8685, Springer, Heidelberg, Germany (2014)
51. Pröll, S., Rauber, A.: Scalable Data Citation in Dynamic, Large Databases: Model and Reference Implementation. In: Hu, X., Young, T.L., Raghavan, V., Wah, B.W., Baeza-Yates, R., Fox, G., Shahabi, C., Smith, M., Yang, Q., Ghani, R., Fan, W., Lempel, R., Nambiar, R. (eds.) Proc. of the 2013 IEEE International Conference on Big Data. pp. 307–312. IEEE (2013)
52. Silvello, G.: Structural and Content Queries on the Nested Sets Model. In: Ferro, N., Tanca, L. (eds.) Proc. 20th Italian Symposium on Advanced Database Systems (SEBD 2012). pp. 283—288. Edizioni Libreria Progetto, Padova, Italy (2012)
53. Silvello, G.: A Methodology for Citing Linked Open Data Subsets. D-Lib Magazine 21(1/2) (2015), http://dx.doi.org/10.1045/january2015-silvello
54. Silvello, G.: Linked open data framework for serendipity in history of art research. In: Ferilli, S., Ferro, N. (eds.) Proc. of 1st AI*IA Workshop on Intelligent Techniques At LIbraries and Archives co-located with XIV Conference of the Italian Association for Artificial Intelligence, IT@LIA@AI*IA 2015. CEUR Workshop Proceedings, vol. 1509. CEUR-WS.org (2015)
55. Silvello, G.: Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data. Journal of the American Society for Information Science and Technology (JASIST) in print, 1–28 (2016)
56. Silvello, G., Bordea, G., Ferro, N., Buitelaar, P., Bogers, T.: Semantic Representation and Enrichment of Information Retrieval Experimental Data. International Journal on Digital Libraries (IJDL) in press, 1–28 (2016)
57. Silvello, G., Ferro, N.: "Data Citation is Coming". Introduction to the Special Issue on Data Citation. Bulletin of IEEE Technical Committee on Digital Libraries (IEEE-TCDL) 12(1), 1–5 (May 2016)
58. Silvello, G., Ferro, N.: "Data Citation is Coming". Introduction to the special issue on data citation. Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation 12(1), 1–5 (May 2016)
59. Steiner, C.M., Agosti, M., Sweetnam, M.S., Hillemann, E.C., Orio, N., Ponchia, C., Hampson, C., Munnelly, G., Nussbaumer, A., Albert, D., Conlan, O.: Evaluating a digital humanities research environment: the CULTURA approach. International Journal on Digital Libraries (IJDL) 15(1), 53–70 (2014)
60. Thanos, C., Borri, F., Candela, L. (eds.): Digital Libraries: Research and Development. First International DELOS Conference. Revised Selected Papers. Lecture Notes in Computer Science (LNCS) 4877, Springer, Heidelberg, Germany (2007)
61. Vernooy-Gerritsen, M.: Enhanced Publications: Linking Publications and Research Data in Digital Repositories. Amsterdam University Press (2009)