Related to other papers in this special issue	4 (p40); 18 (p181)
Addressing FAIR principles	A, I

FAIR Data and Services in Biodiversity Science and Geoscience

Larry Lannom[1†], Dimitris Koureas[2] & Alex R. Hardisty[3]

[1]Corporation for National Research Initiatives (CNRI), Reston, Virginia 20191, USA

[2]Naturalis Biodiversity Center, 2333 CR Leiden, The Netherlands

[3]School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, Cardiff, UK

Keywords: Digital Object Architecture (DOA); FAIR; DiSSCo; Biodiversity; Geoscience

Citation: L. Lannom, D. Koureas & A.R. Hardisty. FAIR data and services in biodiversity science and geoscience. Data Intelligence 2(2020), 122–130. doi: 10.1162/dint_a_00034

ABSTRACT

We examine the intersection of the FAIR principles (Findable, Accessible, Interoperable and Reusable), the challenges and opportunities presented by the aggregation of widely distributed and heterogeneous data about biological and geological specimens, and the use of the Digital Object Architecture (DOA) data model and components as an approach to solving those challenges that offers adherence to the FAIR principles as an integral characteristic. This approach will be prototyped in the Distributed System of Scientific Collections (DiSSCo) project, the pan-European Research Infrastructure which aims to unify over 110 natural science collections across 21 countries. We take each of the FAIR principles, discuss them as requirements in the creation of a seamless virtual collection of bio/geo specimen data, and map those requirements to Digital Object components and facilities such as persistent identification, extended data typing, and the use of an additional level of abstraction to normalize existing heterogeneous data structures. The FAIR principles inform and motivate the work and the DO Architecture provides the technical vision to create the seamless virtual collection vitally needed to address scientific questions of societal importance.

† Corresponding author: Larry Lannom (E-mail: llannom@cnri.reston.va.us, ORCID: 0000-0003-1254-7604).

1. INTRODUCTION

For hundreds of years, scientists have collected and studied plants, animals, rocks, minerals and fossils from our planet. Representing the world's known biological and geological diversity, more than 3 billion physical specimens are housed, organized and cataloged as natural science collections (NSC) in thousands of museums around the world. These represent an unparalleled resource, a scientific infrastructure for discovering and documenting the world's bio- and geo-diversity; its past, present and future and its influence on global challenges in environment and society. Today's systems for exploiting this material, however, are slow, expensive, inefficient and limited [1, 2]. Despite significant existing global domain resources such as the Global Biodiversity Information Facility (GBIF), which aggregate and serve primary biodiversity data that include collections-related elements, the systematic absence of linkages to other data classes, such as DNA sequences, literature, ecosystem and medical/chemical data represent significant impediments to maximizing the impact of NSCs. By creating representations of specimens and collections in cyberspace and treating these assets digitally – "*digital specimen*" and "*digital collections*" – it is possible to persistently link data classes together, enabling seamless unified access to information. Such data-rich "*virtual collections*" offer possibilities for wider, more flexible and meaningful access for a varied range of science and policy applications.

2. CHALLENGES

A true virtual collection resource requires that data it holds must be findable, accessible, interoperable and reusable. These FAIR principles [3] are first principles in building this global scientific asset for natural sciences research; as well being first principles in other, and to the degree possible across, research domains.

The importance of achieving data "FAIRness" – the attribute of data to be findable, accessible, interoperable and reusable – in biodiversity science and geoscience is becoming increasingly clear. As humanity confronts the reality of climate change and all that entails in precedents and outcomes, the need to understand reliably and at a fine level of detail – the variety of life and its environment on the planet is essential to our well-being. What has been lost, what is the current state, and what are the trends? And how do changes affect the balance of ecosystems that mankind depends upon for water, food, health, etc.? While the combined natural science collections and information related to those collections do not provide any easy answers, they do form an invaluable information resource that must be fully exploited toward addressing such questions.

Findable: The first requirement in building digital research infrastructures for bio- and geo-diversity is the ability to find relevant resources, starting with the digital specimens and digital collections that anchor the information facet of infrastructure. Finding resources requires that i) they are uniquely and persistently identified, and ii) their identities are closely bound to enough metadata to discover identifiers when those are unknown. In library and information retrieval terms, this represents the difference between a known item search and a subject search. In network terms, the subject search, run against one or more relevant catalogs or indexes, reveals the identifier(s) and the identifier resolution system reveals the network locations

of the identified resources. The current state of specimen resource identification and associated metadata falls far short of this requirement when considered on a global or regional scale but is not impossible today at an institutional level. Meeting this challenge requires completing the massive digitization effort to create the digital surrogates (including metadata) for the physical specimens, based on agreed identifier and metadata standards and, of course, a consistent effort on the part of collection holders to apply those standards. It should also be noted that, like libraries and archives, the resources of interest in this area are expected, and have already been shown, to be useful over centuries. Mechanisms for finding digital specimens and collections must persist over unimaginable changes in technology.

Accessible: Identifying and locating a digital specimen or other resource is not the same as being able to access and use it. Access may require use of an unknown protocol or special permissions. The current sets of digitized specimens and related information are held in various collection management systems with widely different functionalities and management approaches. A recent survey by Koureas (2018, unpublished) across 115 European natural science museums shows more than 100 commercial and in-house solutions are in use. This legacy must be respected and, even following consolidation and harmonization, each different system must be approached on its own terms. This makes it difficult to create a seamless virtual collection and an approach that aggregates multiple heterogeneous collection management system inputs is needed.

Interoperable: The point of building a virtual collection of distributed data resources is to treat those as a unified scientific asset – to be able to easily find and access data across the combined set, and to be able to re-combine and/or otherwise compute across the data to develop new knowledge and test the old. Digital specimens and related data must be represented in a common manner using known formats and metadata schemes without replacing all that exist now. Metadata must be detailed enough for data to be understood by those who do not own and did not create the data. Meaning acquired from interpreting specimens must be made explicit by using appropriate standard representation schemes, and otherwise semantic differences create substantial barriers to interoperability [4]. Additionally, users should not need to know different methods for working with logically similar but locationally separate parts of the collection. The results of applications and analyses that run across the virtual collections must be trusted and that trust comes only from an understanding that apples are compared to apples and oranges to oranges. Applying algorithms, statistical tests, or other analysis to heterogeneous data without understanding whether the measurements or observations being used represent the same information in the same way can result in reasonable sounding results that are misleading in reality. There will be great value in the envisioned virtual collection provided it is built with care and congruent understanding of what is being assembled.

Reusable: Given that data resources can be found, accessed, and sufficiently well represented and understood to be interoperable, those resources can be reused within individual bio- and geo-diversity domains, across domains, and, with effort and care, across related domains. Information is created by interpreting data to attach meaning, and that exists in a context [5]. A pattern of changes in temperature does not mean much if its context is unknown. Each time data cross a boundary into a new context they must carry their original meaning with them or allow the new context to obtain understanding of that

meaning through another mechanism. This will not always be self-evident and data that carried a clear meaning in one context can be in danger of losing it in a new context, e.g., when combined with data originating elsewhere under different circumstances. Addressing this issue, we should make sure that data are not misrepresented in the new context of seamless unified access to bio- and geo-diversity data, or indeed for any aggregating environment it is one of the more formidable challenges in achieving FAIRness. Technology alone cannot solve this problem, but it can provide tools, approaches and standards that make it possible for people to solve [5].

3. DIGITAL OBJECT ARCHITECTURE

One approach to solving the challenges described above, to aggregating and manipulating heterogeneous research data, is the set of principles embodied in Digital Object Architecture (DOA). DOA began at CNRI, the home institution of Lannom, but as interest and use has grown it has been handed off for the public good to the non-profit DONA Foundation, Geneva, Switzerland. This is from the DONA website [6]:

> *"The Digital Object (DO) Architecture (also known as the DO Architecture or simply the DOA) is a logical extension of the Internet architecture that addresses the need to support information management more generally than just conveying information in digital form from one location in the Internet to another. The DOA enables interoperability across participating information systems, whether in the Internet or not. It is a non-proprietary architecture and is publicly available."*

The approach is described in detail elsewhere [6, 7, 8]. We describe it only briefly here to show how it can be applied to address the challenges outlined above. DOA's fundamental benefit to the management of heterogeneous data is to provide a means of grouping, managing and processing fragments of data and information in a uniform manner through a new layer of abstraction – the digital object. There is a rich history of adding layers of abstraction to solve problems of complexity in computing and information management and the DO Architecture continues this trend. The Internet today is a prime example: it provides a virtual network connecting many heterogeneous networks through use of routers and a single address space (the Internet Protocol (IP) address). Computer operating systems we all use every day provide a layer of abstraction over bit storage using files, thus allowing general interoperability across computing platforms. High-level programming languages combined with interpreters and compilers allow software applications to be used with ease across multiple computing environments. The DO Architecture aims to do this with networked data and information, as Figure 1 illustrates.

Services shown in the "cloud" can be orchestrated to provide an object view of underlying storage, e.g., file systems, or basic data management systems such as databases. The resulting set of identified objects provide a common, and constant, view with "remote control" management of data distributed in various locations and systems, which can change without changing the virtualized object. These services exist today in one form or another. The well-established and successful use of persistent identifiers for scholarly journal articles is one such example. However, others are not yet widely used, and few are tightly coordinated and orchestrated in the way needed.

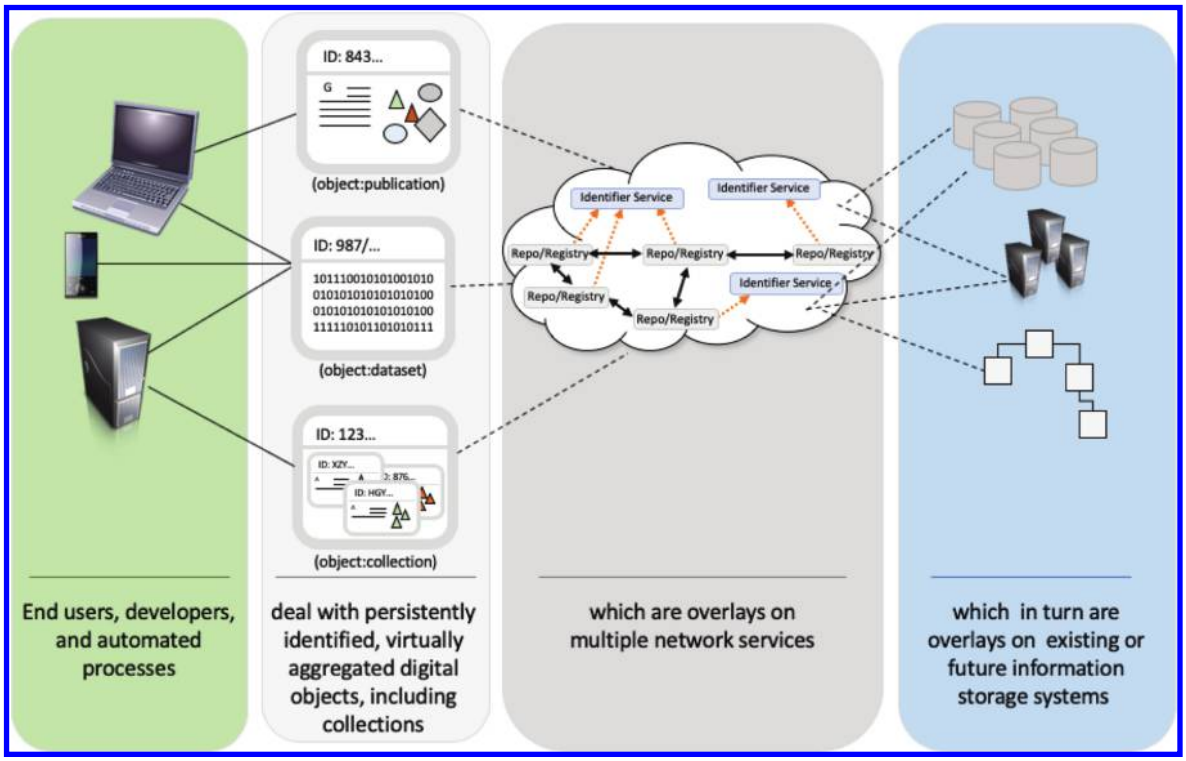

Figure 1. Digital Object Architecture enabling networked data and information interoperability and management across heterogeneous systems.

We do not claim that simply articulating this vision solves the enormous challenges involved in creating seamless distributed research environments, e.g., semantic interoperability [9] but we feel strongly that it provides a way forward that matches well with the FAIR principles [3, 10]. It needs to be developed and evaluated, both for specific disciplines, as is being proposed here, and as an approach to more generic scientific data challenges, such as the findability of provenance data coming out of workflow automation [11].

4. ADDRESSING THE CHALLENGES

Findable: Referential integrity – the ability to reference items reliably and persistently over time and beyond changes in technology – is key to addressing findability. Persistent identifiers are at the heart of the DO Architecture, with each DO assigned an identifier and that identifier globally resolvable to current state data about the DO, e.g., where and how to access data. Note here that everything in the DO Architecture is treated as an object; thus, metadata objects are also uniquely and persistently identified. Metadata can be tightly linked with the object to which it refers and, from the point of view of client software, included with the object, or can be linked through identifiers with each metadata object containing the identifier of

the object it describes as well as its own identifier. Searching across appropriate collections of metadata to find the identifier(s) of relevant object(s) requires aggregating that metadata either centrally ahead of time, in real-time across distributed collections, or some combination of the two. In the DO Architecture this is the role of one or more metadata registries, which are special forms of repositories providing access to metadata objects, of which, the registry maintained for film and television assets by the Entertainment Identifier Registry Association (EIDR) is one example [12].

Building FAIR data and services begins with creating digital objects. Establishing the proper granularity level at which to apply DOA is essential. What exactly receives an identifier and becomes the first-class citizen in the environment? In the case of NSC the answer is clear – the digital surrogate of the physical specimen is the primary object type. From that point on, the metadata (especially provenance of the specimen) can be made part of the object by mapping or brokering underlying information into object form, as illustrated in Figure 1. Metadata is made accessible through a registry presenting uniform access methods across heterogeneous collection management systems.

Accessible: Once identified and found, NSC data must be accessible. The identifier system can help with this, providing authentication and authorization requirements as part of the current state data of the identified object(s). Repositories serve as object portals, regardless of where or how the data are stored. This level of indirection allows clients to use a single access protocol across multiple underlying information organization schemes – a role fulfilled by the Digital Object Interface Protocol (DOIP) [13].

Interoperability: Interoperability is a key challenge presented by heterogeneous scientific data collections and is the raison d'être of the DO approach. Multiple information systems confront the research community with multiple access paths, multiple data organization and representation schemes, varying degrees of metadata completeness and heterogeneous methods. Invoking a digital object approach does not solve these problems by itself but offers a set of approaches to ameliorate difficulties. Mapping multiple schemes into one or a small number of common schemes, identifying those schemes, and associating a set of methods or named operations with each scheme type allows client software to navigate across and operate within multiple environments without detailed knowledge of underlying systems.

A key ingredient of DO Architecture addressing interoperability is an extended notion of data types. This was explored by the Research Data Alliance [14], has been adopted as an ICT standard in public procurement [15] and is currently under consideration by ISO. These data types are intended to serve as an additional level of indirection such that the type of an object can be associated with a set of common characteristics and identified processes and operations, tied together in one or more publicly accessible registries of types.

Reusable: All the principles of the Digital Object Architecture come together in reusability. To leverage existing data, it must be findable, accessible, and interoperable. By simplifying the current level of heterogeneity through an added layer of abstraction, the resulting objects are ready for further investigation and reuse.

The work of building this set of interoperable objects within a domain of biodiversity and geoscience data appears daunting, but the results promise to be worthwhile. When a pattern of creating and linking

objects becomes clear, it can be pursued over time and increasing amounts of data will become available for further research and for linking to new data.

The DO Architecture provides concepts and existing components that can act as a stable basis for FAIR-based research infrastructure for biodiversity and geodiversity science, as illustrated by example in Figure 2 for the Distributed System of Scientific Collections (DiSSCo)[16].

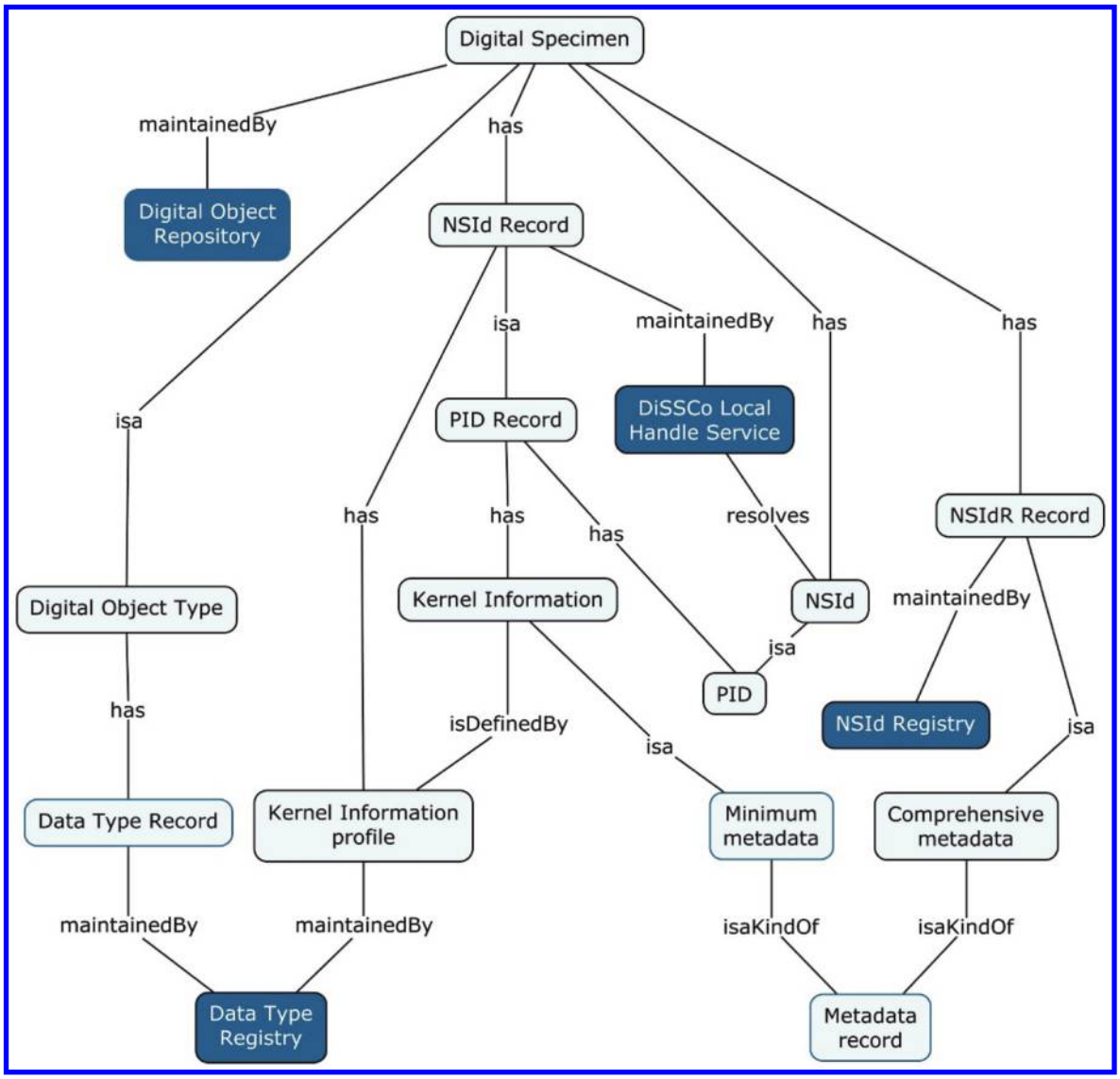

Figure 2. DiSSCo Digital specimen concept (top) and relations to Digital Object Architecture concepts (light blue) and components (dark blue). PID = persistent identifier. NSId = Natural Science Identifier.

5. DISTRIBUTED SYSTEM OF SCIENTIFIC COLLECTIONS (DISSCO)

DiSSCo is a priority pan-European Research Infrastructure aiming to unify more than 110 individual NSCs across 21 countries into a single open and FAIR scientific information resource [17]. DiSSCo evaluates the DO Architecture approach toward building seamless access for bio/geo specimen data, beginning with digitized surrogates of over a billion specimens from natural science collections across Europe. The Handle System [17, 18] will serve as the basis for the Natural Science Identifier (NSId). DiSSCo will prototype use of the CORDRA digital object repository [19]. These, plus other existing components and concepts related to the DO Architecture, including PID Kernel information [20] and data types [14], are shown in context in Figure 2. The FAIR principles inform and motivate the work and the DO Architecture provides the technical vision to create the seamless virtual collection vitally needed to address scientific questions of societal importance.

AUTHOR CONTRIBUTIONS

L. Lannom (llannom@cnri.reston.va.us) is a key member of the team at Corporation for National Research Initiatives (CNRI) that has designed and developed the Digital Object Architecture. A.R. Hardisty (hardistyar@cardiff.ac.uk) and D. Koureas (dimitris.koureas@naturalis.nl) conceived and investigated use of Digital Object Architecture to address challenges in building research infrastructure for biodiversity science and geoscience. All authors contributed equally to the writing, review and approval of the present article.

REFERENCES

[1] R.E. Gropp. Specimens, collections, and tools for future biodiversity-related research. BioScience 68(1)(2018), 3–4. doi: 10.1093/biosci/bix155.

[2] D.E. Schindel & J.A. Cook. The next generation of natural history collections. PLOS Biology 16(7)(2018), e2006125. doi: 10.1371/journal.pbio.2006125.

[3] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

[4] M. Stocker, P. Paasonen, M. Fiebig, M.A. Zaidan & A. Hardisty. Curating scientific information in knowledge infrastructures. Data Science Journal 17(2018), 21. doi: 10.5334/dsj-2018-021.

[5] A. Aamodt & M. Nygård. Different roles and mutual dependencies of data, information, and knowledge—An AI perspective on their integration. Data & Knowledge Engineering 16(3)(1995), 191–222. doi: 10.1016/0169-023x(95)00017-m.

[6] DONA Foundation. (2019). Digital Object Architecture. Available at: https://www.dona.net/digitalobject architecture.

[7] R. Kahn & R. Wilensky. A framework for distributed digital object services. International Journal on Digital Libraries 6(2)(2006), 115–123. doi: 10.1007/s00799-005-0128-x.

[8] P.J. Denning & R.E. Kahn. The long quest for universal information access. Communications of the ACM 53(12)(2010), 34. doi: 10.1145/1859204.1859218.

[9] G. Guizzardi. Ontology, ontologies and the "I" of FAIR. Data Intelligence 2(2020), 181–191. doi: 10.1162/dint_a_00040.
[10] P. Wittenburg, G. Strawn, B. Mons, L. Boninho & E. Schultes. Digital objects as drivers towards convergence in data infrastructures. Technical paper. doi: 10.23728/b2share.b605d85809ca45679b110719b6c6cb11.
[11] T. Weigel, U. Schwardmann, J. Klump, S. Bendoukha & R. Quick. Making data and workflows findable for machines. Data Intelligence 2(2020), 40–46. doi: 10.1162/dint_a_00026.
[12] Entertainment Identifier Registry Association (EIDR). Entertainment identifier registry search page. Available at: https://ui.eidr.org/search.
[13] DONA Foundation. Digital Object Interface Protocol Specification. Version 2 (12 November, 2018). Available at: https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf.
[14] L. Lannom, D. Broeder & G. Manepalli. RDA Data Type Registries Working Group Output (April 30, 2015). doi: 10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458.
[15] European Commission. Commission Implementing Decision (EU) 2017/1358 of 20 July 2017 on the identification of ICT Technical Specifications for referencing in public procurement (Text with EEA relevance). C/2017/5055 OJ L 190, 21.7.2017, pp. 16–19. Available at: http://data.europa.eu/eli/dec_impl/2017/1358/oj.
[16] DiSSCo. Distributed system of scientific collections. Available at: https://dissco.eu/.
[17] Corporation for National Research Initiatives (CNRI). Handle.Net Version 9, Technical Manual. Available at: http://hdl.handle.net/20.1000/113.
[18] DONA Foundation. The Handle System. Available at: https://www.dona.net/handle-system.
[19] Corporation for National Research Initiatives (CNRI). Cordra Digital Object Repository and Registry software. Available at: https://www.cordra.org/.
[20] T. Weigel, B. Plale, M. Parsons, G. Zhou, Y. Luo, U. Schwardmann, … & K. Kurakawa. RDA recommendation on PID kernel information (Version 1). doi: 10.15497/rda00031.

This article has been cited by:

1. Annika Jacobsen, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, Barend Mons, Erik Schultes, Marco Roos, Mark Thompson. 2020. A Generic Workflow for the Data FAIRification Process. *Data Intelligence* **2**:1-2, 56-65. [Abstract] [Full Text] [PDF] [PDF Plus]
2. Barend Mons, Erik Schultes, Fenghong Liu, Annika Jacobsen. 2020. The FAIR Principles: First Generation Implementation Choices and Challenges. *Data Intelligence* **2**:1-2, 1-9. [Citation] [Full Text] [PDF] [PDF Plus]
3. Oya Beyan, Ananya Choudhury, Johan van Soest, Oliver Kohlbacher, Lukas Zimmermann, Holger Stenzhorn, Md. Rezaul Karim, Michel Dumontier, Stefan Decker, Luiz Olavo Bonino da Silva Santos, Andre Dekker. 2020. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. *Data Intelligence* **2**:1-2, 96-107. [Abstract] [Full Text] [PDF] [PDF Plus]