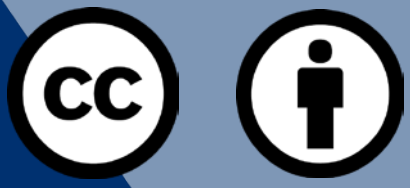
October 2016

Interoperability Standards - Digital Objects in Their Own Right

Susanna-Assunta Sansone and Philippe Rocca-Serra



Cite this as: Sansone, Susanna-Assunta & Rocca-Serra, Philippe (2016) Interoperability Standards - Digital Objects in Their Own Right. Wellcome Trust.
https://dx.doi.org/10.6084/m9.figshare.4055496

Interoperability Standards - Digital Objects in Their Own Right

Susanna-Assunta Sansone and Philippe Rocca-Serra

1. Executive Summary

Interoperability standards enable the operational processes underlying exchange and sharing of information between different systems to ensure all digital research outputs are Findable, Accessible, Interoperable and Reusable, according to the FAIR principles[1]. Among the interoperability standards, one category focuses on the descriptions (or metadata) of digital objects. Within this category there are content standards, which opens datasets to transparent interpretation, verification and exchange.

The uptake of content standards is vital for high-quality, reproducible research and for the integrative analysis and comparison of heterogeneous data from multiple sources, domains and disciplines. When a content standard is mature and appropriate standard-compliant systems become available, these must then be channelled to the appropriate stakeholder community, who in turn must recommend them (in data policies) or use them to facilitate a high-quality data cycle, from data generation to standardization, and through to publication and subsequent sharing and reuse.

Providing an introduction to the landscape of standards, this report focuses on the wealth of content standards available in the life and biomedical sciences, introducing their life cycle and the ecosystem of stakeholders and initiatives. The report also highlights a number of technical, social a financial pain points, which must be addressed if we are to realize a vision of integrable content standards that seamlessly become part of the research and data management enterprise of the future.

Over and above all, we have two recommendations.

1. **As Data Science culture grows, digital research outputs (such as data, computational analysis and software) are being established as first-class citizens. This cultural shift is required to go one step further: to recognize interoperability standards as digital objects in their own right, with their associated research, development and educational activities.**
2. **New funding frameworks need to be created to provide catalytic support for activities necessary to: research new or apply existing methods to develop, extend, refine and harmonize interoperability standards, and also related tools and educational material. Launch joint funding frameworks and/or match fundings opportunities - among relevant funding agencies - on specific domains, within and cross-disciplines.**

[1] Wilkinson MD et al. “The FAIR Guiding Principles for scientific data management and stewardship”. *Sci Data*. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18

2. Introducing Standards

Standards are agreed-upon conventions for doing something, e.g. managing a process or delivering a service, and are established by community consensus or an authority. In this report the focus is on the specifications, guidelines or criteria designed to ensure data and any other digital object (such as code, algorithms, workflows, models, software, or journal articles) are FAIR. Defined and endorsed by a growing community, these principles put a specific emphasis on enhancing the ability of machines to automatically find and use digital objects, in addition to supporting its reuse by individuals throughout their life cycle.

There are several types of standards set to ensure all digital objects are FAIR; standards themselves are digital assets. There is no agreement on how to best categorize them, these standards address different, but often complementary needs, that include but are not limited to:

- Machine-processable descriptions - e.g., minimum reporting requirements, terminologies, file formats or conceptual models for citation, credit and interoperability purpose.
- Identification for discovery, citation and credit;
- Accessibility of the information - e.g., access permission, data protection, patient consent, anonymization and encryption;
- Indicators or metrics to measure performance, use and quality;
- Versioning and documentation practices - e.g., for code, algorithms or tools;
- Tracking provenance of and relationships between digital concepts - e.g., interpretations and conclusions;
- Analysis - e.g., standardized descriptions of the workflow and related software used.

The perspective and focus of certain standards varies, ranging from standards with a specific biological or clinical domain of study (e.g., stem cells, clinical trials, neuroscience) or significance (e.g., to model and predict biological processes), to the technology used to generate the dataset (e.g., imaging modality, high-throughput sequencing). The motivation for the development of these standards spans from the creation of *de novo* standards (e.g., to fill an existing gap), to the mapping and harmonization of complementary or contrasting efforts, or the extension and repurposing of existing standards.

2.1. Stakeholders Community

Standardization activities are numerous and diverse, driven by large organizations with industrial strength or taking place at a grass root level. Profiling the parties involved gives an insight into the complexity of the standards life cycle and related challenges. There is an extensive range of stakeholders involved in these efforts, illustrating the large number of players involved in the data life cycle. These include: domain experts (usually leading researchers in the area, but also data producers and clinicians), technical experts (e.g., ontology engineers, knowledge engineers, data architects, software developers), data managers (establishing data management plans, or supporting researchers), quality officers (who are responsible for ensuring procedures are adhered to and data produced meet the expected grade), government

officials, policy makers, librarians, data scientists, curators, trainers, software and lab equipment vendors, service providers, journal publishers and funders.

Stakeholders are involved in managing, serving, curating, preserving, publishing or regulating data and/or other digital objects; they are often - but not always - not only producers but also end users of standards. Each stakeholder or stakeholder group is interested in different aspects of a particular standard and play different roles (e.g., provide use cases to inform the development of standards, implement in tools, endorse in policies) in different phases of its life cycle. In an ideal scenario, for example, standards should be implemented by dedicated experts in tools, services and infrastructure and made 'invisible' to the lay users of these systems who often have little or no familiarity with standards.

2.2. Interoperability Standards

Life and biomedical sciences increasingly require effective ways to find, access and (re)use data and related code or software (e.g., for computational aggregation, integration and comparison). Interoperability standards enable the operational processes, underlying exchange and sharing of information between different systems. Optimal interoperability is achieved when access and use of data and other digital objects is completely automated, and accessible to both human and machine. This requires standardized: (i) identifiers and (ii) descriptions (or metadata) for each digital object, including the accessibility level of the information and/or licence type. Identifiers and metadata would then need to be implemented by an array of registries, catalogues, databases and services that are needed to find, store, manage (e.g., mint, track provenance, version) and aggregate (e.g., interlink and map etc.) these digital objects.

Identifiers are outside the scope of this report. However, it is worth mentioning that there are several type of identifiers[e.g.2,3,4]; unique, resolvable and versionable identifiers are essential elements of the digital word, and common guidance to design new or maintain existing identifiers is extremely important and being addressed[e.g.5].

Machine actionable as well as human consumable metadata standards for digital objects serves for different purposes: to enhance their discoverability, citation, credit as well as their interoperability with related objects, and evaluation for reuse and reproducibility by third parties. According to their purpose, the type, depth and breadth of metadata standards may vary. For example, reproducibility would require a richer metadata than discoverability, citation and credit. Consequently, in the metadata standards space there are several efforts: some are driven by one specific purpose, others meet several needs. The following paragraphs provide exemplars of community-driven metadata standards efforts, focused on one or more digital objects. These

[2] https://www.force11.org/group/resource-identification-initiative
[3] https://permid.org
[4] https://schema.datacite.org
[5] McMurry J et al. "10 Simple rules for design, provision, and reuse of identifiers for web-based life science data" doi:10.5281/zenodo.31765

also illustrate the value of synergies between more specific life and biomedical sciences and generic cross-domain metadata efforts.

Infrastructure and metadata standards for software lag substantially behind that of other digital objects; this is also well documented in the NIH BD2K workshop's report on software discovery[6]. Existing mechanisms used by software repositories, languages and in scientific domains are heterogeneous and there is not a common standard[7]. The CodeMeta[8] effort brings together leaders of software and data repositories with academic researchers to develop a 'crosswalk table' that would translate the diverse metadata currently used. This effort intersects and works with related initiatives, including (but not limited to) the Force11 Software Citation working group[9], the SSI[10] and WSSSPE[11].

To enhance discoverability (e.g., by search engines), aggregations (e.g., by data indices) and analysis of content in different websites and services, it is essential that the metadata served by these resources is consistently structured. Many groups operate in these areas. Under the W3C[12] organization - where members work together with the public to develop open standards and technology stack to support the semantic web - several groups focus on the health care life science; one group in particular works to define the essential but broadly applicable metadata elements for a dataset description[13]. Another initiative is Bioschemas[14], which works (i) to encourage the use of the schema.org[15], a structured semantic markup for web pages' content used by the main search engines; and (ii) to coordinate its extension, where needed, in the life science area. The Bioschemas' WGs focus on different digital objects, including tools, training material and datasets, but also organization, events and more, bringing together members from a variety of communities, including (but not limited to) ELIXIR[16] and the ELIXIR-UK Node[17], Pistoia Alliance[18], Goblet[19], BioSharing[20], BBMRI[21] and the EMBL Australia Bioinformatics Resource[22].

---

[6] "Software Discovery workshop" NIH BD2K, May 2014: http://www.softwarediscoveryindex.org

[7] Howison, J. and Bullard, J. (2015), Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *J Assn Inf Sci Tec*. doi:10.1002/asi.23538

[8] http://codemeta.github.io

[9] https://www.force11.org/software-citation-principles

[10] https://www.software.ac.uk

[11] http://wssspe.researchcomputing.org.uk

[12] https://www.w3.org

[13] https://www.w3.org/TR/hcls-dataset

[14] http://bioschemas.org

[15] http://schema.org

[16] https://www.elixir-europe.org

[17] http://elixir-uk.org

[18] http://www.pistoiaalliance.org

[19] http://www.mygoblet.org

[20] McQuilton P, et al., "BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences". Database (Oxford). 2016 May 17, doi: 10.1093/database/baw075

[21] http://bbmri-eric.eu

[22] https://www.embl-abr.org.au

The data-related activities in Bioschemas are done also in synergy with the NIH BD2K bioCADDIE project[23]. This is a community-driven effort creating DataMed[24], a data discovery index, which does for data what PubMed[25] has done for the literature. The prototype development is driven by a set of queries (or competency questions), elicited from researchers, to search and interrogate the datasets. This is essential that the metadata model, powering DataMed, has the necessary breadth and depth of information and structure to answer the queries. This underlying metadata model, named DATS (DatA Tag Suite), follows the successful example of JATS (Journal Article Tag Suite)[26], the metadata standard required by PubMed to index the literature. The DATS model has generic core and extended elements, to progressively accommodate domain-specific metadata for more specialized data types, as needed. The model is available as machine readable schemata, annotated using the schema.org semantic markup (a collaboration with the Bioschemas initiative); this will ensure that the metadata index by DataMed benefits from an increased visibility (by search engines and tools), increased accessibility (via common query interfaces), and possibly, an improve in search ranking. Currently, work is in progress to ensure DataMed harvests (DATS-formatted) metadata of (biomedically relevant) datasets stored in a variety of repositories, and from other metadata aggregators. Examples of the later include: the HeartBD2K OmicsDI[27], indexing metadata of transcriptomics, genomics, proteomics and metabolomics datasets; and DataCite, indexing metadata of DOI-identifier datasets. The NIH BD2K bioCADDIE also funds a series of pilots on and around data discovery and harvesting, to complement the DataMed prototype; one pilot focuses on data citation. The Force11 Data Citation Implementation Group[28], a set of diverse stakeholders and organizations (including DataCite and CODATA) behind the Joint Declaration of Data Citation Principles[29], has agreed to a set of minimal requirements for repositories to implement a landing page with metadata supporting data citation.

In addition to these above-mentioned efforts, in the life, environmental and biomedical sciences there is a wealth of standardization efforts focussing on deeper and domain specific metadata to maximize reusability, reproducibility, interpretation and verification of datasets. Known also as content standards, these cover the what, who, when, where, how and why. The following section focus entirely on this type of metadata standards.

3. Focus on Content Standards

Data comparability and reproducibility is still a big issue[e.g,30], and reusability of datasets is a central aspect of data preservation. Content standards encompass all elements of a dataset,

[23] https://biocaddie.org
[24] https://datamed.org
[25] http://www.ncbi.nlm.nih.gov/pubmed
[26] https://jats.nlm.nih.gov/index.html
[27] http://www.omicsdi.org
[28] https://www.force11.org/group/dcip
[29] Starr, J. et al. “Achieving human and machine accessibility of cited data in scholarly publications”. *PeerJ Comput. Sci.* 1, e1. 2015 doi: 10.7717/peerj-cs.1
[30] Mobley, A. "A Survey on Data Reproducibility in Cancer Research ..." 2013. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0063221>

allowing the fundamental biological entities (e.g., samples, genes, cells), experimental components (e.g., conditions, cell lines), complex concepts (such as bioprocesses, tissues and diseases), the analytical process and the mathematical models and their instantiation in computational simulations (spanning from the molecular level through to whole populations of individuals) to be harmonized with respect to structure, format and annotation. Hereafter these elements are collectively referred as datasets.

By ensuring the information is reported consistently, efficiently and meaningfully, content standards open datasets to transparent interpretation, verification, exchange, reuse, integrative analysis and comparison. Community-driven content standards such as MIAME[31] and GO[32] have become essential resources in modern molecular biology and computational biology. The value of and adherence to content standards is recognized and recommended by an increasing number of reports, concordats and policies in and around open research data[e.g.33,34,35,36,37].

3.1. Mapping the Landscape

Although it is generally agreed that the use of open, community-developed standards is critical, far less agreed upon is exactly which data standards should be used, the criteria by which one should choose a standard, or even what constitutes a data standard[38]. While very few community-developed content standards are known in other disciplines, as listed by the JISC DCC directory[39], over a thousand exist in the life, environmental and biomedical sciences. In these areas BioSharing[40] is building a comprehensive curated resource that maps this landscape. As an informative resource, BioSharing ensures that standards are findable and accessible (according to the FAIR principles). As an educational resource, BioSharing works to provide the indicators necessary to monitor the development, evolution and integration of standards. By interlinking[41] standards, databases and data policies (from funders, journals and other organizations), BioSharing guides users to discover those standards that are implemented

[31] Brazma, A. et al. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data." *Nature genetics* 29.4 (2001): 365-371.
[32] Ashburner, Michael et al. "Gene Ontology: tool for the unification of biology." *Nature genetics* 25.1 (2000): 25-29.
[33] NIH Data Sharing Policy: http://grants.nih.gov/grants/policy/data_sharing
[34] Wellcome Trust: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTVM050569.pdf
[35] Concordat on open research data, Higher Education Funding Council for England (HEFCE), Research Councils UK (RCUK), Universities UK (UUK) and Wellcome Trust: http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf
[36] Alan Turing Institute “Symposium on reproducibility for data-intensive research” report 21.07.2016. Figshare. https://dx.doi.org/10.6084/m9.figshare.3487382.v2
[37] Boulton G. et al, “Science as an open enterprise” 2012. [Online]. Available: https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf
[38] Tenenbaum JD, Sansone SA, Haendel M. “A sea of standards for omics data: sink or swim?”. J Am Med Inform Assoc. 2014 Mar-Apr;21(2):200-3. doi: 10.1136/amiajnl-2013-002066.
[39] http://www.dcc.ac.uk/resources/metadata-standards
[40] https://biosharing.org
[41] https://biosharing.org/search/?q=

by databases, and to find the policies that refer to them[42], providing evidence of use and other important indicators that users take into consideration when selecting a resource. Working with and for researchers, developers, curators, funders, journal editors, librarians and data managers, BioSharing helps producers of standards (databases and policies) to ensure their resources are findable by prospective users, and enable consumers to make an informed decision as to which standard (database or policy) to (re)use or endorse. Operating since 2011, BioSharing is driven by an international advisory board, operates as an open WG under Force11 and the RDA[43], collaborates with NIH BD2K and EMBL Australia Bioinformatics Resource among others , and has recently become a ELIXIR-UK Node resource[44].

With over a thousand content standards and thousands of databases, curating and interlinking their descriptions is a lengthy process, especially as both the coverage and status of these resources must be verified with their respective communities, and in many cases require frequent updates. The BioSharing team and their collaborators are working to paint the complete picture of this dynamic and complex landscape of content standards, linking to/importing from other complementary and specialized portals. Although the work is far from complete, preliminary insights are summarized in the following paragraphs.

There is no agreement on how to best name the types of content standards; below these are broadly divided into several categories, and their known total number is given as of 30th of August 2016.

- *Reporting guidelines* or *checklists* outline the necessary and sufficient (or minimum) information that is vital for contextualizing and understanding a dataset. These vary from general guidance to itemised prescriptions of the information that should be provided. There are 106 reporting guidelines in BioSharing[45] (which supersedes the MIBBI portal[46], and complements the 322 medical-focused reporting guidelines currently in the EQUATOR Network[47], which seeks to improve the reliability and value of published health research literature by promoting transparent and accurate reporting).
- *Models/formats* or *syntaxes* define the structure and interrelation of information from a conceptual model or schema, and the transmission format, such as XML, CSV or RDF, to facilitate data exchange between different systems. There are 204 models/formats in BioSharing[48].
- *Terminology artefacts* or *semantics* provide the unambiguous identification and definition of concepts within a scientific domain. These add an interpretive layer to the information beyond any that might be provided by the syntax, and enable complex grouping and querying of the data. Encompassing lexicon, dictionary, vocabularies, taxonomies,

[42] https://biosharing.org/recommendations
[43] https://rd-alliance.org/group/biosharing-registry-connecting-data-policies-standards-databases-life-sciences.html
[44] http://elixir-uk.org/node-resources-1
[45] https://biosharing.org/standards/?q=&selected_facets=type_exact:reporting%20guideline
[46] Taylor C., et al. “Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project” *Nature Biotechnology* 26, 889 - 896 (2008) doi:10.1038/nbt.1411
[47] http://www.equator-network.org
[48] https://biosharing.org/standards/?q=&selected_facets=type_exact:model/format

thesauri and ontologies, there are 345 terminology artefacts in BioSharing[49]. These are also linked to the respective records marked as 'production' in BioPortal, containing a total 531 terminology artefacts.

- *Common Data Elements (CDEs)* are particularly used in clinical research, patient registries, and other human subject research in order to improve data quality and opportunities for comparison and combination of data from multiple studies and with electronic health records. There are over 18351 elements in the NIH CDEs Repository[50].

3.2. Key Organizations and Stakeholders

There is an extensive range of communities involved in these standardization efforts, as described in section 2.1. Standard organizations have different level of formality (e.g., some are legal entities, while the majority are ad hoc working groups), membership types (e.g., open and free, members only), operational approaches (e.g., organized in formal committee or as open working groups) and funding levels. Awareness of which organizations are doing what in which specific domain is vital to a coordinated approach, especially to minimize overlapping, fragmented and competing alternatives (see section 3.4); since the activities of these group changes over time, such awareness must be continually updated. Categorizing the organizations is a not trivial and will not be attempted in these following paragraphs, but key exemplars are also presented to illustrate their diversity and relations, in some cases.

There are two main drivers of content standard generation, and consequently, two types of outputs. Those developed by formal Standards Developing Organizations (SDOs), such as HL7[51] (health) and CDISC[52] (clinical) are known as *de jure* standards, as they are prescribed by an official or formal authority. The standards development process is not always open to all interested parties. For example, ISO[53] (who develop both generic[54] and biotechnology-specific standards[55]) is a network of national bodies that - according to their membership type[56] - have a different level of access and influence over their standards work. Generally SDOs sell or license the standards; in the best scenario a subset of standards is licensed at no cost[e.g.57], or if standards are open, charges apply to advanced training or programmatic access to the standards[e.g.58]. Grass-root, bottom-up efforts, such as HUPO-PSI[59] (proteomics), GSC[60]

[49] https://biosharing.org/standards/?q=&selected_facets=type_exact:terminology%20artifact
[50] https://cde.nlm.nih.gov/cde/search
[51] http://www.hl7.org
[52] http://www.cdisc.org
[53] http://www.iso.org
[54] "ISO 8601 - Time and date format." 2013: http://www.iso.org/iso/home/standards/iso8601.htm
[55] http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=4514241
[56] http://www.iso.org/iso/about/iso_members.htm
[57] http://www.hl7.org/implement/standards/nocost.cfm
[58] https://www.cdisc.org/standards
[59] http://www.psidev.info
[60] http://gensc.org

(genomics), Metabolomics Society[61], OME[62] (imaging), TDWG[63] (biodiversity) etc. develop open *de facto* standards, such as SBML[64] and MIABE[65], which are generally directly adopted by the community. The division between content standards by SDOs and grass-roots is one of the major issues, and the topic of section 3.5.

An example of a standards generating network is COMBINE[66], which brings together grass-roots communities to develop content standards for computational modelling. The OBO Foundry[67] is an example of an umbrella organization for ontologies; it brings together groups who are committed to adhering to common development principles that ensure ontologies are orthogonal and interoperable. There is also a number of alliances bringing together leading organizations in diverse sectors to lower barriers and accelerate development and scientific progression. Exemplars are the GA4GH[68] initiative, which works to create a common framework of harmonized approaches to enable the responsible, voluntary, and secure sharing of genomic and clinical data. And the Pistoia Alliance, which is a group of life sciences industry experts who use pre-competitive collaboration to address issues around aggregating, accessing, and sharing data that are essential to innovation in R&D; the ontology mapping project[69] is an exemplar activity. In addition, there are cross-disciplinary, multidisciplinary and transdisciplinary efforts, such as the RDA[70] (open data sharing), Force11 (modern scholarly communications), CODATA (data quality, reliability, management and accessibility) and JISC[71] (UK digital services and solutions). Bridging across these diverse efforts, within life and biomedical areas and across domains, is a major challenge.

As producers and consumers of content standards, initiatives that develop data and knowledge infrastructures are themselves major stakeholders. For example, the Innovative Medicine Initiative, Europe's largest public-private initiative, has clearly identified content standards as an essential components of translational research[72], funding FAIR-enabling infrastructure projects such as OpenPHACTS[73] (pharmacology) and eTRIKS[74] (translational medicine). The latter

[61] http://metabolomicssociety.org/board/scientific-task-groups/data-standards-task-group
[62] http://www.openmicroscopy.org
[63] http://www.tdwg.org
[64] Hucka M., et al. “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models” Bioinformatics (2003) 19 (4): 524-531. doi:10.1093/bioinformatics/btg015
[65] Orchard S., et al., “Minimum information about a bioactive entity (MIABE)”. Nature Reviews Drug Discovery 10, (2011). doi:10.1038/nrd3503
[66] http://co.mbine.org
[67] Smith et, al., “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration” *Nature Biotechnology* 25, 1251 - 1255 (2007) doi:10.1038/nbt1346
[68] http://genomicsandhealth.org
[69] http://www.pistoiaalliance.org/projects/ontologies-mapping
[70] https://rd-alliance.org
[71] https://www.jisc.ac.uk
[72] Martin A, et al. “Data Standards are needed to move Translational Medicine forward” (2012) Transl Med 3.2:: 2161-1025 http://dx.doi.org/10.4172/2161-1025.1000119
[73] https://www.openphacts.org
[74] https://www.etriks.org

project has delivered the "Standards Starter Pack"[75], a guideline providing scientists, project managers and data custodians with a comprehensive overview of the content standards in the clinical and functional genomics areas. The list of recommended standards is also available as a collection in BioSharing[76].

The FAIR-supporting ELIXIR is set to orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments, creating an infrastructure to integrate research data from all corners of Europe and ensure a seamless service provision that is easily accessible to all. ELIXIR's first implementation project, EXCELERATE, has an interoperability backbone[77] at the core, focussing on identifiers, content standards (also in collaboration with BioSharing) and related operational practices and services. This interoperability activity is run in coordination with CORBEL[78], which brings ELIXIR together other biological and medical research ESFRI infrastructures sharing the same data management principles[79]: BBMRI[80] (biobanking), EATRIS[81] (translational research), ECRIN[82] (clinical trial), InfraFrontier[83] (functional genomics), INSTRUCT[84] (structural biology), EU-OpenScreen[85] (chemical biology), EMBRC[86] (marine organism), EuroBioImaging[87] (imaging), MIRRI[88] (microorganism) and ISBE[89] (system biology, including FAIRDOM, a component to establish integrated FAIRer data and model management service[90]).

The trans-NIH BD2K initiative supports the research and development of innovative and transforming approaches and tools to enable biomedical research as a digital research enterprise[91]. The BD2K envisions the creation of the *Commons*[92], a shared virtual space for FAIR digital objects, including interoperability standards. To define the award mechanisms, administrative procedures, policies, eligibility requirements, review criteria etc., to best fund community-driven standards organizations, the BD2K initiative ran two workshops in 2013 and 2015: to gain the perspectives of both large, longstanding standards organizations as well as smaller, more loosely organized groups in the basic and clinical sciences, on what has and what

[75] IMI eTRIKS Standards Starter Pack - Release 1.1 April 2016 dx.doi.org/10.5281/zenodo.50825
[76] https://biosharing.org/collection/eTRIKS
[77] https://www.elixir-europe.org/excelerate/interoperability
[78] http://www.corbel-project.eu/home.html
[79] "Principles of data management and sharing at European research infrastructures". Joint working paper by ELIXIR and EU-OPENSCREEN with AnaEE, BBMRI, EATRIS, ECRIN, ERINHA, EMBRC, Euro-BioImaging, INFRAFRONTIER, INSTRUCT, ISBE, LifeWatch and MIRRI doi:10.5281/zenodo.8304
[80] The Biobanking and Biomolecular Resources Research Infrastructure: http://bbmri- eric.eu
[81] The research infrastructure for translational medicine : www.eatris.eu
[82] The European Clinical Research Infrastructure Network: http://www.ecrin.org
[83] The infrastructure for mouse disease models and phenotype data: http://www.infrafrontier.eu
[84] The integrated structural biology unlocking the secrets of life: http://www.structuralbiology.eu
[85] The European Infrastructure of Open Screening Platforms for Chemical Biology: http://www.eu-openscreen.de
[86] The European Marine Biological Resource Centre: http://www.embrc.eu
[87] The research infrastructure for imaging technologies: http://www.eurobioimaging.eu
[88] The microbial resource research infrastructure: http://www.mirri.org
[89] The Infrastructure for Systems Biology in Europe: http://project.isbe.eu
[90] http://fair-dom.org
[91] https://datascience.nih.gov/bd2k
[92] https://datascience.nih.gov/commons

has not worked in these endeavours. The outcome of the first workshop provides information on the standards' life cycle (see section 3.3), the second, an insight in the technical, social and financial pain points (see section 3.4). Currently BD2K-funded centres, such as LINCS[93] (network-based cellular signatures) and HeartBD2K[94] (cardiovascular medicine) are both producers and consumers of content standards. The bioCADDIE's DataMed (described in section 2.2) aims to tag those datasets harvested from standards-compliant databases. Another centre, CEDAR[95], has been funded specifically to develop methods, tools and practices to make authoring complete datasets smarter and faster; automatic generation of descriptive templates will also explore an 'invisible use' of content standards from BioSharing.

Leveraging on its networked membership, and in collaboration with NIH BD2K and ELIXIR-EXCELERATE, BioSharing has recently conducted a survey[96] to gather users' views on which information and functionality a registry of standards should have to help them make informed decisions, e.g., how to best select standards and understand their maturity, or to find the databases that implement them. This 10 question survey has gathered 533 responses from researchers, standard developers, database curators and industry scientists to librarians, funders and journal editors, from all over the world, with a predictable concentration in Europe and the USA. The results[97] show that the information and functionality BioSharing currently provides fulfils approx. 80% of the users needs, and that approx. 65% of the respondents are already familiar with it. Those unmet requirements will drive BioSharing future activities, which are set to cross-link standards (databases and policies) to other digital objects and information portals, including ELIXIR-related one for training material (TeSS)[98], tools and services[99].

3.3. Life Cycle and Indicators

Like any other digital object, standards in general and content standard more specifically have a life cycle. The first NIH BD2K workshop[100] on community-driven content standards provides an invaluable insight on different issues pertain to each phase of the life cycle (*i.e.*, formulation, development and maintenance), showing that communities' social and technical approaches to common problems are also quite diverse. A summary of the workshop's key findings are provided in the next paragraphs.

Formulation is about the identification of a need (e.g., data exchange or reporting), collection of use cases (valuable for defining the breadth and depth of the requirements), definition of the

[93] http://www.lincsproject.org/centers/bd2k-lincs-dcic
[94] http://www.heartbd2k.org
[95] Musen MA et al., "The center for expanded data annotation and retrieval". J Am Med Inform Assoc. 2015 Nov;22(6):1148-52. doi: 10.1093/jamia/ocv048.
[96] https://bd2kccc.org/2016/01/15/biosharing-standards-registry-survey
[97] 10.6084/m9.figshare.3795810
[98] https://tess.elixir-uk.org
[99] Ison J et al. "Tools and data services registry: a community effort to document bioinformatics" resources. Nucleic Acids Res. 2016 Jan;44(D1) D38-47. doi:10.1093/nar/gkv1116.
[100] "Report on Frameworks for Community-Based Standards Effort Workshop" NIH BD2K, Sep 2013: https://datascience.nih.gov/sites/default/files/bd2k/docs/frameworks_for_comm_based_standards_efforts_report.pdf (summary), 10.6084/m9.figshare.3795816 (full report).

scope (what the content standard is supposed to address and what not), and prioritization of the work. The need for a particular content standards effort is typically driven by the needs of the community of practice being made known through any of a number of different channels. For example, direct observation of problems in the research community is employed by the HUPO-PSI, while organized discussions including community polls are the approaches taken by the Metabolomics Society and the GSC. The use of a standing body to which requests for various standards efforts are made is exemplified by the DDI Alliance[101], operating in the social sciences, whose Technical Committee field requests for new standards as well as modifications to existing standards. This formulation phase crucially depends upon identifying, assembling and engaging with the right people (self-appointed, or solicited for particular expertise or role) with iterations punctuated by consultation with experts and the broader community.

Development encompasses iterations of the work (usually by a core group), solicitation of feedback (on the various drafts) and requests for testing and evaluating the work. Although some groups have successfully started the discussion with virtual interactions, face-to-face meetings seem crucial to galvanize a new group, enable participants to evaluate their commitment, design the initial workplan, harmonize different perspectives and explore available options such as whether existing content standard(s) could be reused, modified, or extended for use to meet the identified need. The type and frequency of group interactions depend on the type of content standards developed, the granularity, coverage, and number of people actively involved; e.g., reporting guidelines (narrative or in list form) are less demanding than highly structured models/formats or terminology artefacts. Once the core group roughly shapes the content standards, additional stakeholders are engaged and the effort iterates forward with their multi-pronged input. For example, the various working groups in CDISC therapeutic areas publish roadmaps and calendar updates of their progress[102].

Maintenance is about the creation of exemplar implementations (in tools and/or databases), technical and documentation (to guide wider uptake and further implementations) and education materials; this phase also addresses sustainability, evolution of the content standards, including backward compatibility of each version, via migrations or conversion modules. Typically, content standards efforts are long-term endeavours and require updating and evolution as the science (in the laboratory but also in computational biology) and technology to which they relate change, and when knowledge is obtained, e.g., the classification of a new genetic disorder. Responsibility for keeping efforts updated over the long term varies across groups, from the responsibility of a core group as in many of the grass-roots efforts, to committees elected by the members that follow very formalized processes, such as in HL7.

Although the ultimate indicator of progress and success is wide adoption or extension of a standard, there is almost always a significant lag between the development of the effort and those final outcomes. The last two phases can be especially time consuming and difficult from social, technical and financial perspectives (see section 3.4). Nevertheless, the ability to solicit

[101] http://www.ddialliance.org

[102] "Coalition For Accelerating Standards and Therapies (CFAST)" http://blogs.fda.gov/fdavoice/index.php/tag/coalition-for-accelerating-standards-and-therapies-cfast

testing, monitor results, and manage feedback and requests for extensions are key intermediate milestones for assessing the progress of this phase.

Long lifetime and the sustainability of content standards are best supported by their wide acceptance and adoption, and the continued participation of new groups. Once a content standard become part of the fabric of research, all stakeholders have an incentive to aid its continuance, especially those with commercial interests (e.g., instrument manufacturers or publishers). Succession is also an important part of sustainability. For example, the community behind OBI[103], an ontology for the description of life-science and clinical investigations, has operated since 2004; several groups have contributed at different phases, driven by a core of long standing contributors present throughout its life cycle. This OBO Foundry-compliant ontology also illustrate how it is possible to create an artefact by importing parts of other orthogonal biomedical ontologies such as GO, ChEBI[104], hence reusing without altering their meaning.

Extensions and diversity of these applications (for both research and production purposes) can be important an element to ensure the evolution and long lifetime support of content standards; derivative products are then maintained by an expanded community of developers and users. An example is provided by a FAIR-supporting ISA effort[105], a general-purpose metadata tracking framework customizable for different reporting requirements, terminologies and formats, supported by a set of tools. Recently approved as a resource of the ELIXIR-UK Node, this grass-roots initiative has run since 2007, evolving from an earlier effort initiated under the FGED society (formerly MGED). Currently ISA metadata framework is implemented in many domains[106] and different ways: (i) used as is; (ii) reused as a generic core module for more specialized content standards by others initiatives, e.g., for metabolomics datasets[107]; (iii) extended to fit specific needs, e.g., for characterization nanomaterial (ISA-Tab-Nano[108], a grass-root work also formalized by an SDO[109]) and for plant[110]; (iv) embedded into and extended by other, such as those underpinning the BD2K CEDAR and BD2K DataMed's DATS models; (v) combined to other representation models, like RO[111] and nanopublications[112] to capture the

[103] Bandrowski A et al,. "The Ontology for Biomedical Investigations". PLoS One. 2016 Apr 29;11(4):e0154556. http://dx.doi.org/10.1371/journal.pone.0154556
[104] Hastings, J.. et al., "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013". Nucl. Acids Res. (2013) 41 (D1). doi: 10.1093/nar/gks1146
[105] Sansone et al., "Toward interoperable bioscience data" *Nature Genetics* 44, 121–126 (2012) doi:10.1038/ng.1054
[106] http://www.isacommons.org
[107] Rocca-Serra P., et al. "Data standards can boost metabolomics research, and if there is a will, there is a way" Metabolomics. 2016; 12: 14. doi: 10.1007/s11306-015-0879-3
[108] Baker NA., et al., "Standardizing data" Nat Nanotechnol 8(2):73-4. 2013 doi: 10.1038/nnano.2013.12
[109] https://www.astm.org/Standards/E2909.htm
[110] http://cropnet.pl/phenotypes
[111] Belhajjame K. et al., "Using a suite of ontologies for preserving workflow-centric research objects" Web Semantics: Science, Services and Agents on the World Wide Web 2015. doi:10.1016/j.websem.2015.01.003
[112] Mons B. et al., " The value of data". Nature Genetics. 201143(4):281–283 http://dx.doi.org/10.1038/ng0411-281

experimental processes and reproduce the findings in a scientific paper[113]; and lastly (vi) generalized to describe and discover datasets associated to publications, such as in the Springer Nature *Scientific Data* journal[114,115].

Tracking the life cycle of content standards efforts is very challenging task. Two recent domain-specific surveys give a unique insight in data sharing practices and use of content standards. In the system biology community, the results[116] show that - compared to earlier findings[117] - the number of reporting guidelines has grown, along with the availability and uptake of formats, and the latter are the most widely used type of standard. In the clinical metabolomics community, results[118] show awareness of technology-focused standards (as those by the Metabolomics Society and HUPO-PSI) but low penetration of the clinical standards (by SDOs, like HL7 and CDISC) especially in the academic community, an issue expanded in section 3.5.

Grass-roots standardization efforts can cease to operate, or become dormant for a period: primarily due to the lack of funds or a succession plan, or because the community has accomplished its mission and the interest of the driving players has moved to others areas. From a personal communication (to board members) it seems that the FGED society[119] - the long standing grass-roots initiative that developed MIAME - is in the process of closing its operation. Funded in 1999 as MGED, the society focused on content standards for DNA microarray experiments and in 2010 changed its name to reflect its enlarged scope of activities. Despite the society closing down, its content standards are implemented by several communities in annotation tools and databases, which will look after their maintenance.

BioSharing works to paint the complete picture of this dynamic and complex landscape of efforts and players. As the community within each effort is the most reliable source of information, BioSharing also crowdsources information to update and curate the description and status of each standard. Currently four indicators are used to tag each content standard (and database), providing information about its readiness for implementation or use.

- **R:** ready for use, implementation, or recommendation. The majority of standards have this tag[120].
- **Dev**: in development. Perhaps a standard is being actively developed but isn't quite ready for use[e.g.121].

---

[113] Gonzalez-Beltran A., et al., "From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing: The Complementary Roles of Data Models and Workflows in Bioinformatics." 2015 PLoS One 10(7):e0127612. doi: 10.1371/journal.pone.0127612

[114] http://scientificdata.isa-explorer.org

[115] http://www.nature.com/sdata

[116] Stanford NJ et al., "The evolution of standards and data management practices in systems biology". Mol Syst Biol. 2015 Dec 23;11(12):851. doi: 10.15252/msb.20156053

[117] Klipp et al., "System biology standards - the community speaks". Nat Biotechnol 25 (2007). http://dx.doi.org/10.1038/nbt0407-390

[118] Personal communication from authors; publication in process: https://blogs.biomedcentral.com/gigablog/2016/07/19/guest-posting-building-phenomenal-metabolomics-e-infrastructure

[119] http://fged.org

[120] https://biosharing.org/standards/?q=&selected_facets=status:Ready

[121] https://biosharing.org/bsg-s000642

- **U**: uncertain. When we are unsure as to whether a standard is in development, active or deprecated, and attempts to reach out to the developing community has failed[e.g.122]
- **D**: deprecated. When it is known and confirmed that a standard is no longer maintained or active; if known, a note is added to give the reason for the deprecation, e.g., subsumed[e.g.123], or superseded[e.g.124] and in these cases a link to the extant record(s) are also provided.

3.4. Challenges for Producers and Consumers

The general mobilization of grass-roots groups and SDOs, and the growing number of data policies by funders, publishers and other organizations, are tangible and positive signs of the movement for open and reproducible research. Despite this wealth of initiatives, the production and use of content standards still remain challenging practices[125] due a number of technical, social and financial pain points that were also highlighted by the first and second NIH BD2K workshops on community-driven content standards[126]. There are several issues one needs to be aware of, both as a producer and consumer of content standards, especially as the roles are intertwined and interchangeable.

There is **no central authority** for standards, or at least one that is recognized by all the parties involved, to coordinate the development of orthogonal and integrable efforts; this has led to **overlapping** and **competing** alternative standards, **some are open some not**. A key **separation** exists between SDOs and grass-roots initiatives, and especially between the research, and clinical and medicine sectors, which are regularized and have their own set of standards (more in section 3.5). Several **synergistic activities** work to foster harmonisation and consolidation of open content standards: within one type (e.g., OBO Foundry: ontology), a domain (e.g, GA4GH: clinical genomics), a discipline (e.g., NIH BD2K: biomedical), or cross-discipline (e.g., RDA). As the Royal Society report on "Science as an open enterprise" highlights, the drive for broad standards should not override the specific needs of disciplinary and domain communities. The latter are regarded as important because they address 'real world' requirements of content standards; e.g., for a particular technology being used or the particular biologically or medically-delineated community concerned. However, remaining bounded by a particular discipline or domain has the unfortunate consequence that these standardisation efforts in general remain **fragmented**, leading to the development of (arbitrarily) different content standards, thereby limiting their combined used. For example, data producers of datasets in which source material has been subject to several kinds of analyses (e.g., genomic sequencing, protein-protein interaction assays and the clinical measurement) find

[122] https://biosharing.org/bsg-s000047
[123] https://biosharing.org/bsg-s000583
[124] https://biosharing.org/bsg-s000005
[125] Sansone SA and Rocca-Serra P. "On the evolving portfolio of community-standards and data sharing policies: turning challenges into new opportunities". Gigascience. 2012;1(1):10. doi:10.1186/2047-217X-1-10
[126] "Executive Summary - Workshop on Community-based Data and Metadata Standards Development: Best practices to support health y development and maximize impact" NIH BD2K, Feb 2015: https://datascience.nih.gov/sites/default/files/bd2k/docs/ExecSumm_CBDMSworkshopFEB2015.pdf

particularly challenging to share datasets as coherent units of research because of the diversity of content standards with which the parts must be formally represented. Researchers, acting as data consumers, need to understand the various reporting guidelines, terminologies and models/formats used to reassemble these fragmented datasets scattered across databases. Ideally, **standards should stand alone but** should **also function well together**, especially to better support multi-dimensional investigations but also the aggregation of pre-existing datasets from one or more domains.

The vision is for integrable content standards that become part of the research process may just remain a wish, unless the **common pain points** (affecting individual as well as synergistic initiatives) are recognized and addressed. From a **technical perspective** it is necessary to remove unnecessary **duplications** between the domains that are covered by existing content standards, but also to identify **gaps** and foster initiatives in these new areas that are not covered. An **operational infrastructure** and **governance framework** is essential (yet not always formalized) to specify how development or harmonisation of standards should be achieved, e.g., to handle conflicts, updates and versions, how the extensive work program can be subdivided amongst the involved parties. Lack of **tools** and **services** around standards is a major bottleneck; these are few and scattered, but essential elements to facilitate quicker and more widespread adoption. Requirements include components to track requests or modifications to standards, querying and managing (programmatic) access[e.g.127] to them, validation[e.g.128] (compliance to) and to convert (between/among) standards[e.g.129] along with lookup and mapping services[e.g.130] and annotation/curation applications[e.g.131,132,133] to reduce the time, knowledge and skills required to facilitate use and buy-in.

Faced with a dearth of efforts, consumers of standards may not always be equipped to navigate, select or recommend the most appropriate standards and often see them as burdensome and/or over-prescriptive, especially in the absence of tools and services that facilitate their 'invisible use'. Few **training materials** and **events** exist for the development and use of standards, in particular for those developed by grass roots initiatives. A **portal**, like BioSharing, is an essential element to explore the landscape of initiatives, to discover standards e.g., by their domain of coverage and readiness for use, or based on implementations (in databases and tools) and recommendations (in data policies) and to also understand the relationships between content standards (e.g., which model/format fulfills which reporting guideline) and evolution (e.g., which standard is superseded by which).

Although quite difficult, these technical and documentation hurdles are not insurmountable. In contrast, the **sociological barriers** involved in these kinds of large-scale, multi-stakeholders endeavours can be far more challenging. Extensive **community liaison** and communication

[127] http://www.cdisc.org/standards/share
[128] http://www.psidev.info/validator
[129] https://github.com/ISA-tools/isa-api
[130] http://www.ebi.ac.uk/ols/index
[131] https://www.ebi.ac.uk/fg/annotare
[132] http://www.rightfield.org.uk
[133] http://isa-tools.org

need to be managed; invaluable feedback cycles need to be recorded, and the complex **stakeholders' dynamics** unpacked. **Incentives** and **rewards** need to be identified for all contributors, producers and prospective consumers. Managing the technical and social components throughout the standards life cycle, takes time, resources, and expertise. Ownership of open content standards can also be problematic; the **legal framework** to encourage their maintenance, contribution and evolution is very embryonic. This is especially relevant when in addition to industry, commercial content suppliers, aggregators and publishers are involved. In many instances, these organizations are supportive of open grass-root standards; in others they have developed their proprietary or even open standards.

Last but not least, is the **cost** of overcoming these challenges: these demanding tasks require specific **funding frameworks.** SDOs rely on revenues from memberships, subscriptions, licences and training events, even if often the material is free to nonprofit and regulatory authorities[134]. **Sustainability** is a major challenge for the majority of grass-root efforts run with the contributions of volunteers, whose only reward is often co-authorship in publications. Funding for grass-root standardization initiatives has traditionally been limited and relied on a small number of individuals or relatively short-term grants, or travel funds for face-to-face meetings. In other cases these activities form (often small) components in research grants. In the best case scenario standardization activities are core elements of infrastructure projects.

3.5. Bridging the Divide: Basic Research, Clinical and Medicine worlds

Successful reuse (e.g., integration) of clinical and basic research data requires significant laborious, often manual intervention to match up and identify all digital entities of interest, such as molecules, compounds, cells, observations, drugs etc. Reconstructing and understanding the humane physiome, for example, is a highly demanding computational challenge requiring multiple disparate data sources, a number of tools and a variety of standards[e.g.135]. Another example is provided by the food, health, medical and life science industries, which to inform and enhance the decision-making process, has invested heavily in people, procedures and tools that integrate internal datasets with publicly available research data, information commercially produced (licensed data and knowledge bases) and outsourced (to contract research organizations)[e.g.136]. These are major challenges, exacerbated by the fact that, whilst some (not all) basic research data is annotated using grass-roots standards, health information are commonly structured and defined by purpose specific models[e.g. 137]and clinical

[134] http://www.meddra.org/subscription/subscription-rate
[135] Nickerson et al., "The Human Physiome: how standards, software and innovative service infrastructures are providing the building blocks to make it achievable". Interface Focus February 19, 2016 http://dx.doi.org/10.1098/rsfs.2015.0103
[136] D. Searls. "Data integration: challenges for drug discovery". Nat. Rev. Drug Discov., 4 (2005). http://dx.doi.org/10.1038/nrd1608
[137] http://www.openehr.org

terminologies[e.g.138], and/or follow the standards requirements set by regulatory bodies, such those by the FDA[139] and EMEA[140].

These two categories of standards are not interoperable, and a common set of open content standards accepted by all stakeholders does not exist; many have highlighted the value of addressing this challenge via pre-competitive initiatives[e.g.141,142], such as IMI. A group of researchers from academia and industry - brought together by the Pistoia Alliance - illustrates the issue[143]. Human disease data is an good example of this problem, being defined by a variety of non-interoperable and diverse terminologies, including as ICD[144], MeSH[145], NCI Thesaurus[146] and HDO[147]. The consequence of this interoperability issue is a painstaking mapping exercise between these terminologies (and/or other type of content standards used to structure their data), resulting in a **combinatorial explosion of cross-referencing** required to align the (same) entity across each data source. This **mapping work is duplicated** by each industry **internally and never opening shared**, with notably rare exceptions[e.g.148]. Furthermore, these terminologies are often developed for specific purposes, and are unable to support different applications: e.g., a thesaurus is useful for text-mining, but may be poor for classification tasks. Only **pre-competitively developed open** content standards - available to all information producers and consumers - are able to assist industry (and the wider community) in the efficient management, processing and application of internal and external data, which are so vital to R&D productivity. This will save costs, reduce redundancy, ensure greater coverage and a wider body of expertise. Given the vast landscape of biomedicine and therefore content standards, the authors also suggest that it should be possible to identify many major areas of common need, such as open cell/tissue hierarchies and inter-relationships, catalogs of animal models (and relationships to human biology), pathophysiological processes and disease phenotypes.

In addition to many of the challenges identified in section 3.4, the creation of pre-competitive open content standards that harmonize or bridge between SDOs and grass-roots products present additional challenges. This would also require a **legal framework** to deal with technical or legal restrictions that could prevent the cross-referencing of proprietary vocabularies.

---

138 http://www.ihtsdo.org/snomed-ct

139 http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm

140 http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000645.jsp&mid=WC0b01ac058078fbe2

141 M.R. Barnes, *et al.* "Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery". Nat. Rev. Drug Discov. (2009). doi: 10.1038/nrd2944

142 Sidders B.et al., "Precompetitive activity to address the biological data needs of drug discovery"Nature Reviews Drug Discovery 13, 83–84 (2014) doi:10.1038/nrd4230

143 Harland et al., "Empowering industrial research with shared biomedical vocabularies" Drug Discov Today. 2011 16(21-22). doi: 10.1016/j.drudis.2011.09.013

144 http://www.who.int/classifications/icd/en

145 https://www.nlm.nih.gov/mesh

146 https://ncit.nci.nih.gov/ncitbrowser

147 http://disease-ontology.org

148 http://pp1.eppo.org

Additionally, **rules of engagement** to manage the extensive community liaison, with rewards and incentives identified for all contributors, whether from the commercial or public sector, are necessary. Industry, however, cannot be the sole mechanism for **funding** this work.

4. Turning Evidence into Recommendations

Technical and social pain points and the exemplars described in the sections above should be used for future target actions. Below, is a list of key needs, with some indicated as priority.

- Need to recognize **standards as digital objects**, coupling standards to effective research data management, and **professionalizing** the role of **the scientists** dedicated to their research, development and implementation.
- Need to create a **dedicated funding framework.**
- Need for a **portal** for discovery of standards, **mapping the landscape** - tracking evolution and status - to reveal the existing coverage (and lack thereof) in different life, biomedical domains and disciplines.
- Need for formal **indicators** and **evaluation** methods to measure standards usage and usability.
- Need for both **incentives** and **credit/recognition** mechanisms.
- Need to enable the development of open standards, maximizing **reuse, modification, extension and integration** of existing standards, **filling gaps** with **new activities** in those domains where standards do not exist.
- Need to foster the **global/worldwide** collaboration and **harmonization** of standards, within each type, but also within and across domains and disciplines
- Needs for open-source **infrastructures, tools and services** to overcome technical and social challenges throughout the standards life cycle, also enabling their use in the data management process.
- Need for greater **coordination** between the activities in **basic research**, **clinical** and **medicine** worlds, via pre-competitive initiatives.
- Need to foster collaboration **beyond the pharmaceutical and biotech industries**, including others key stakeholders such as **publishers**, **librarians**.
- Need for **education**, **documentation**, **hackathons**, **training and courses** materials (and **events**) targeting both producers and consumers of standards, and set to create a new **career path**.
- Need for **business models** to tackle **sustainability**.

The Royal Society “Science as an open enterprise” report, which also highlights the value of standards, recommends that the costs of preparing data and metadata for curation should be included as part of the costs of the research process. This, however, assumes the existence

of an array of standards ready for use as part of creation and delivery of a **data management plan**. As this report illustrates, this is far from the current situation in the life and biomedical sciences. In the vast majority of cases, the development and use of standards are perceived as a service that is 'automagically' executed at little or no cost. Contrarily, there are: **expertise**, **knowledge** and **skills** that must be **professionalized**, such as that happening for biocuration[149], in order to conduct proper **research**, **development** and educational **activities** in and around the development, harmonization and use of standards.

Along with the **recognition of interoperability standards as digital objects in their own right**, comes the need for covering the associated **costs** via dedicated **funding frameworks**. Whilst it is mainly the use of content and other interoperability standards that is funded as an element of e-infrastructure projects, very few funding programmes and calls also support the development of standards[e.g.150]. This status quo will not change dramatically without app funds for (i) the appropriate full-time personnel (to be both recruited and trained) who are dedicated to perform the technical work and manage the social aspect; (ii) the person-hour contributions by several experts, stakeholders, implementers (to be recognized and covered); (iii) core infrastructures, tools and services (to be developed and delivered) to support the life cycle; (iv) and the hackathons, training events, documentation and dissemination (to be organized, run and produced).

An **exemplar** to follow is the recent **NIH BD2K funding opportunity**[151] that builds on the outcomes of its two workshops on standards. This new opportunity is explicitly dedicated to provide "time-limited, catalytic support for activities necessary to develop or extend/refine data and metadata standards and/or related tools in areas relevant to the NIH basic, translational, and clinical research mission". The call supports activities at any point in the standards lifecycle; encourages building on existing partnerships, infrastructure, and resources (whenever possible); and requires the results to be made freely available, and standards deposited in BioSharing (and terminology in BioPortal, which the first is interlinked to). The application budgets are limited to $250,000 direct costs per year, maximum project period is up to 3 years.

To ensure standards are truly global, and that the geographically-distributed stakeholders are engaged, the ideal scenario would be the creation of **joint funding frameworks** and/or **match fundings** opportunities - among relevant funding agencies - on **specific domains, within and cross-disciplines**. New funding frameworks are essential to provide catalytic support for activities necessary to: research new or apply existing methods to develop, extend, refine and harmonize interoperability standards, but also related tools and educational material.

[149] http://biocuration.org
[150] http://www.bbsrc.ac.uk/funding/filter/2016-bioinformatics-biological-resources-fund
[151] http://grants.nih.gov/grants/guide/rfa-files/RFA-ES-16-010.html

October 2016

Version 1

Wellcome exists to improve health for everyone by helping great ideas to thrive. We're a global charitable foundation, both politically and financially independent. We support scientists and researchers, take on big problems, fuel imaginations and spark debate.

**Wellcome Trust, 215 Euston Road,
London NW1 2BE, UK
T +44 (0)20 7611 8888, F +44 (0)20 7611 8545,
E contact@wellcome.ac.uk, wellcome.ac.uk**