A choice of persistent identifier schemes for the Distributed System of Scientific Collections (DiSSCo)

Alex R Hardisty[‡], Wouter Addink[§,|,¶], Falko Glöckler[#], Anton Güntsch[¤], Sharif Islam[§,|], Claus Weiland[«]

‡ Cardiff University, Cardiff, United Kingdom
§ Naturalis Biodiversity Center, Leiden, Netherlands
| Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
¶ Species 2000 Secretariat, Leiden, Netherlands
# Museum für Naturkunde Berlin, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany
¤ Freie Universität Berlin, Berlin, Germany
« Senckenberg - Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main, Germany

Corresponding author: Alex R Hardisty (hardistyar@cardiff.ac.uk)

Academic editor: Editorial Secretary

Abstract

Persistent identifiers (PID) to identify digital representations of physical specimens in natural science collections (i.e., digital specimens) unambiguously and uniquely on the Internet are one of the mechanisms for digitally transforming collections-based science. Digital Specimen PIDs contribute to building and maintaining long-term community trust in the accuracy and authenticity of the scientific data to be managed and presented by the Distributed System of Scientific Collections (DiSSCo) research infrastructure planned in Europe to commence implementation in 2024. Not only are such PIDs valid over the very long timescales common in the heritage sector but they can also transcend changes in underlying technologies of their implementation. They are part of the mechanism for widening access to natural science collections. DiSSCo technical experts previously selected the Handle System as the choice to meet core PID requirements.

Using a two-step approach, this options appraisal captures, characterises and analyses different alternative Handle-based PID schemes and the possible operational modes of use. In a first step a weighting and ranking the options has been applied followed by a structured qualitative assessment of social and technical compliance across several assessment dimensions: levels of scalability, community trust, persistence, governance, appropriateness of the scheme and suitability for future global adoption. The results are discussed in relation to branding, community perceptions and global context to determine a preferred PID scheme for DiSSCo that also has potential for adoption and acceptance globally.

DiSSCo will adopt a ‘driven-by DOI’ persistent identifier (PID) scheme customised with natural sciences community characteristics. Establishing a new Registration Agency in collaboration with the International DOI Foundation is a practical way forward to support the FAIR (findable, accessible interoperable, reusable) data architecture of DiSSCo research infrastructure. This approach is compatible with the policies of the European Open Science Cloud (EOSC) and is aligned to existing practices across the global community of natural science collections.

Keywords

Distributed System of Scientific Collections, DiSSCo, persistent identifier, pid, persistent identifier scheme, pid scheme options

Introduction

“*Digital science means a radical transformation of the nature of science and innovation due to the integration of ICT in the research process and the Internet culture of openness and sharing*.” Source: European Commission (2013).

A radical transformation across Europe for how collections of physical specimens are accessed, harnessed and exploited for scientific research and innovation is the mission for a collections-based digital science in which increased use of information and communication technologies (ICT) allows natural science collections to empower global society with reliable knowledge and evidence to solve natural world challenges. By exploiting data locked, for example in the preserved specimens of herbarium collections, machine learning models can lead to new approaches that can project future climate change (Pearson et al. 2020) and accelerate conservation action (Albani Rocchetti et al. 2021). Such specimens in biological and geological collections that represent our planet's diversity are the curated hard evidence base that underpins important economic activity, environmental and health protection and other policymaking to address troubling societal challenges of the 21st century (green energy, loss of biodiversity, climate change, food/ water security, infectious diseases, etc.). In many cases, they provide the only available source of data on species distributions, geological events and climate in the past.

The Distributed System of Scientific Collections (DiSSCo) is a new Research Infrastructure (RI) of European natural science collections planned to commence full operations in 2026. Novel services for indexing, enriching and assisting reuse of specimens' data, which are expected to begin in pilot form in 2022 dictate that DiSSCo must soon achieve technical and organisational readiness for operating a persistent identification scheme. Durable persistent identifiers (PID) are needed for the extended Digital Specimen digital objects (Webster 2017, BCoN 2019) derived from the physical objects in the natural science collections of participating institutions (Hardisty 2019, Hardisty et al. 2020). The aim of the present options appraisal is to compare options and to choose and present the best strategy to reach the goal to operate a persistent identification scheme, keeping in mind

that the choice should have potential for global adoption across the entire scientific collections community.

Strategic context

DiSSCo represents the largest ever formal agreement (120+ institutions across 21 countries) between natural history museums, botanical gardens and collection-holding universities in Europe. With strategic pillars that cover digitizing and mobilizing content from collections; harmonising data policies, processes and workflows; and maximising the use of expertise, enhancing skills and engaging communities, DiSSCo looks towards new models of data management and infrastructure to achieve its ambitions. With adherence to the FAIR Guiding Principles as an integral characteristic (Lannom et al. 2020) DiSSCo's chosen ICT model is Digital Object Architecture (Kahn and Wilensky 2006) in which Digital Specimens are digital representations on the Internet that act as complementary surrogates for physical specimens in natural science collections.

We use the term 'Digital Specimen' throughout the present article, noting at the time of writing/publication that a worldwide discussion under the umbrella of the Alliance for Biodiversity Knowledge is taking place. This discussion*[1] on technical convergence of DiSSCo's Digital Specimen concept as explained below and the very similar concept emerging from the Extended Specimen Network strategy of the Biodiversity Collections Network (BCoN) in the USA (BCoN 2019, Lendemer et al. 2019) is expected to reach consensus on the new term 'Digital Extended Specimen' (DES) circumscribing the Digital Specimen and Extended Specimen ideas in one technical concept. Nevertheless, the terms can be used interchangeably and interpreted to mean the same in the future.

Persistent identifiers and FAIR for open digital science

In the 21st century, science is increasingly assisted by computers and other digital machines like remote sensors, digitization pipelines and genome sequencers. In this digital science persistent identifiers (PIDs) play a role to identify people (researchers, collectors, technicians), their organizations and the things they work with (their research inputs and outputs - data, software, literature - and artefacts such as instruments and other physical objects). In an easy-to-read introduction, Meadows et al. (2019) explain the role of PIDs and their value as foundational elements in the overall research information infrastructure. Critically, PIDs act not only as identifiers but also as connectors – of one identified thing to another where a relation can optionally be expressed – 'A is-employed-by B', 'C was-published-by A', 'D was-sequenced-by E', etc. Such connections can be followed both by machines and by humans, not only revealing and providing access to a wide corpus of associated information, but also allowing relations to be understood, insights to be gained and conclusions to be reached. Importantly, the mechanism helps with attribution and recognition of personal and organisational scientific achievements. PIDs make it easier to evaluate the impact of publicly funded work. As noted by Meadows et al. (2019), "*DOIs for publications are a great example of how this works in practice*". But they also note that

extending the use and support of persistent identifiers further requires stronger community commitments.

One such commitment that leads to *de facto* wider use of PIDs is commitment to the FAIR Guiding Principles (Wilkinson et al. 2016, Mons et al. 2017) on making data Findable, Accessible, Interoperable, and Reusable (FAIR). Principle F1 states that both data and metadata should be assigned a globally unique and eternally persistent identifier, while principle A1 states that data and metadata should be retrievable via their identifier. Adopted by the European Commission (EC) as a pillar of their policy objectives to advance the global open science movement, FAIR has become one of the strategic priorities for developing the European Open Science Cloud (EOSC) as well as an essential mandate for all European research infrastructures (European Commission 2018a, European Commission 2018b, European Commission 2019a, European Commission 2019b, European Commission 2019c, Schouppe and Burgelman 2018). Meeting this EC mandate (i.e., exhibiting 'FAIRness' as a characteristic) occurs when a research infrastructure such as DiSSCo achieves and maintains '*toto genere'* *[2]compliance with all fifteen of the FAIR Guiding Principles.

As a strategic priority, persistent identifiers are at the core of the EOSC interoperability framework (European Commission 2021b) with a policy governing their use (European Commission 2020). Also mentioned is the "*the atomic entity for a FAIR ecosystem*", the FAIR Digital Object (European Commission 2018b). These FAIR Digital Objects, as explained in detail by De Smedt et al. (2020) are self-contained, typed, machine-actionable data packages unambiguously identified with persistent identifiers.

By choosing Digital Object Architecture (Hardisty et al. 2020) DiSSCo achieves FAIRness by default and aligns to the main European science-research infrastructure policies that recognise the value and impact of both persistent identifiers and FAIR Digital Objects i.e., Report on turning FAIR into reality (European Commission 2018b), the EOSC implementation plan (European Commission 2019a), and the EOSC interoperability framework (European Commission 2021b). Everything is identified; thus, everything is findable. Everything is described, and these descriptions (metadata), including how to access things are themselves identified. Persons (users), organisations, workflows and machines are identified. Thus, access can be controlled based on identity. Specific typed definitions of 'FAIR Digital Object' (again, each uniquely identified to avoid ambiguity) such as Digital Specimens, Digital Collections*[3], Loans and Visits Transactions, Annotations constrain the heterogeneity and incompatibility of data types across systems, thus contributing to interoperability and ensuring the possibility of bidirectional and interdisciplinary interactions between communities such as our natural sciences community and others. By leveraging the first principles and simplifying complexity using persistent identifiers and definitions of FAIR Digital Object types, the resulting objects are available for reuse (Lannom et al. 2020).

Digital specimens and persistent identifiers

A Digital Specimen is a specific kind of FAIR Digital Object that acts on the Internet as a digital surrogate for a physical specimen in a collection. A Digital Specimen, referenced by its unique persistent identifier represents the sum of digital information on the Internet about a physical specimen object in a natural sciences collection and other data derived from or associated with that specimen. A PID for a Digital Specimen complements identifiers of the physical specimens themselves and/or their corresponding digital database records in institutional collection management systems. Such identifiers include: CETAF Stable Identifiers (Güntsch et al. 2017), International Geo Sample Numbers (IGSN) (Lehnert et al. 2019), GUIDs, Darwin Core Triplets, institution/collection codes and catalog numbers.

As a new kind of object alongside natural objects and fabricated tools (Hui 2012), machine-actionable Digital Specimen objects on the Internet are amenable to processing by and transport between heterogeneous information systems. Such digital objects are editable, interactive, reprogrammable, and distributable (Kallinikos et al. 2013). Interoperability difficulties are much reduced by the definition mechanisms for object types and operations that underlie the concept. Such objects have the implicit capability to remain findable, accessible, interoperable and reusable (FAIR) over timescales familiar to collection-holding institutions. This is many decades (100+ years). As a kind of FAIR Digital Object, Digital Specimens assist to integrate collection data in the data rich world of (*inter alia*) the Earth System Sciences and Life Sciences.

Persistent identifiers that unambiguously identify Digital Specimens (Fig. 1) are one mechanism for digitally transforming collections-based science*[4]. They are are integral elements of the modern-day open science value chain. Through their sustained persistence PIDs contribute to building and maintaining long-term community trust in the accuracy and authenticity of the scientific data being managed and presented. Not only are they valid over the very long timescales common in the heritage sector but they are also capable of transcending changes in underlying technologies of their implementation. They are foreseen as one of the mechanisms for widening access to natural science collections, for transforming collections-based science into the digital era and for helping this community to fully embark on open science practices.

Persistent identifiers make it possible to reliably refer to and find the digital equivalent of a specific specimen held in the collection of a specific institution, and to reliably access data derived from that specimen. This is especially important in the case of voucher specimens, which are the representative samples of identified organisms providing the verifiable evidence for authenticating taxonomic identification (Culley 2013). As the example in Suppl. material 1 shows, community ability to effectively link digital representations of voucher specimens with other data types, such as literature, people, genetic sequence information, traits, or even to assert and sustain semantic links between vouchers continues to be seriously hindered by the lack of PIDs and related services.

Persistent identifiers for specimens on the Internet can be compared in importance with digital object identifiers (DOI) for journal articles and datasets, where an article and its associated supplementary data is referenced via a unique DOI®⋆5. Within scholarly publishing DOIs under the governance of the (International) DOI Foundation are having a transformational effect. Academic journal articles are more accessible and third-party services (such as those of Crossref) based on the growing PID graph are becoming more widely available and used (Cousijn et al. 2021). It is, for example, increasingly possible to link research outputs to the project grants and funding initiatives under which those were produced, as well as to work out the relations and collaboration among authors. In the future it will become possible to associate downstream impacts with the research outputs that created them, thus enabling measurements of value for taxpayers' money. Similar effects are being seen in other industry sectors, such as film/tv and construction where DOIs are being adopted.

In collections-based science, an interconnected specimen graph based on PIDs can, in the future will reveal connections between specimens, and between specimens and the data derived from them or associated with them. An innovative range of exploratory, analytical and mining services is likely to emerge to exploit the graph, leading increasingly to industrialisation of research with machine-actionable metadata allowing for automated actions on data and flexible cross-domain data discovery and recombination.

The Handle System

The Handle System (Sun et al. 2003) is a PID scheme that is used widely, in many guises. As the basis for a persistent identifier mechanism for DiSSCo it is a consequence of the choice of the Digital Object Architecture as the architectural basis of DiSSCo data infrastructure (Hardisty et al. 2020). Digital objects of all kinds, including Digital Specimens need to be identified and they need a global resolution mechanism. The Handle System provides both of these things, whilst also being capable of both exhibiting and exploiting loose coupling with Web technologies.

The Handle System has several characteristics that are very desirable from the DiSSCo perspective:

- The namestrings used for the identifiers are location independent, making them immune to changes in the location of the identified object (provided the binding between the name and the location of the object is properly maintained);
- As pointers to objects, Handles are easily repairable when they break;
- There is transparency for identifying objects of any kind - both physical and digital;
- Identifiers are impervious to changes in underlying implementation technology over the curation timescales typical of the natural science collections sector (many decades, with a target of 100+ years); and,
- The Global Handle Registry infrastructure has reliability, robustness, and resilience for continuous distributed Handle resolution.

Thus, the Handle System and its multiple implementation options is the focus of the present appraisal.

The choice is further justified because FAIR Digital Objects (of which Digital Specimens are a kind) and their Handle-based PIDs are core elements of the European Open Science Cloud. Achieving compatibility with identifiers within the European Open Science Cloud (EOSC) for interdisciplinary interoperability and reusability is a further important benefit of investing in support for a DiSSCo PID mechanism.

Requirements of a Handle-based PID scheme

DiSSCo has six principle requirements when it comes to evaluating alternative Handle-based PID schemes. Explained in the following subsections, these are:

- Scalability
- Trust
- Persistence
- Governance
- Use of appropriate identifiers
- Potential for global adoption.

Scalability

Requirement: Scale for specimens, scale for machines, scale for performance, scale for global use.

Estimates suggest there are approximately 3 billion specimens in natural science collections around the world (Duckworth et al. 1993, Wheeler et al. 2012). Half of those are in Europe, housed in 120+ collection-holding institutions across more than 21 countries. These are used daily by more than 5,000 full-time scientists. Estimates suggest approximately 300 million specimens have been digitized to some degree to date in Europe, although not all that data is yet shared and publicly available.

When digitized, each resulting 'Digital Specimen' must be persistently and unambiguously identified. Subsequent events or transactions associated with the Digital Specimen, such as annotation and/or modification by a scientist must be recorded, stored and also unambiguously identified.

Suppl. material 2 provides an estimate of the number of PIDs likely to be needed throughout the DiSSCo lifetime and beyond. The numbers of identifiers that will be needed will rapidly surpass (by two orders of magnitude) the hundreds of millions of persistent identifiers already in use today. Thus, the scalability of a PID scheme to potentially handle several hundred billion identifiers is a very important requirement. At such levels, and with such large numbers of identifiers, machine-assisted services for working with PIDs will be essential.

A further important consideration is the way control can be exerted over the administration and use of the parts of the namespace hierarchy allocated for DiSSCo. Control and delegation of control over a large range of second or third-level name segments is likely to be needed. By subdividing control for administration and resolution purposes into smaller name segment blocks, better (i.e., faster, more balanced) performance can be achieved. High throughput workflows with multiple Digital Specimens as inputs and actionable steps that involve machine learning algorithms can rapidly generate tens of thousands of artifacts, each of which might need identifying; not forgetting that such workflows also require rapid resolution responses when manipulating intermediate artifacts.

Trust

Requirement: User confidence in the PID scheme, seeing it as appropriate to their needs and trustworthy.

Building trust doesn't happen overnight or by accident. Trust is a result of specific actions that address consistency, quality and excellence. Being consistent with a PID scheme is about reliably and continuously sustaining services for creating, assigning and resolving PIDs over the long-term. A quality PID scheme (in the sense of a scheme that conforms to all the requirements stakeholders place on it) unambiguously delivers accurate and authentic data, traceable back to its points of origin and described by metadata adjusted to the specific community needs. Excellence in a PID scheme comes from being able to operate the scheme responsively, efficiently and cost effectively in a way that encourages convergence towards common digital scientific curation, publishing and access/use policies and practices.

Consistent, high-quality and excellent delivery of PID services built behind a branding that creates instant recognition for PIDs of the chosen scheme as the unambiguous pointers to specific accurate and authentic digital data about a specimen, including an unbreakable link to the corresponding physical specimen acts to confer authority. Trustworthiness should follow. Information quality is the strongest factor to influence organizational benefits through perceived usefulness and user satisfaction (Park et al. 2011).

Persistence

Requirement: Many decades, more than 100 years.

An appropriate PID scheme must address the very long-term needs for durable identification in natural science collections, which can extend over many decades to 100 years or more. Specifically, the resolution component must be capable of accurately resolving PIDs and redirecting users and machines to the location of the data, even after new PIDs cease to be created and assigned. A maintenance responsibility and function for that must continue until such time as PIDs are no longer used for referencing digital specimen data. It's clear that persistence goes hand in glove with governance.

Governance

Governance: An internationally acceptable governance mechanism by stakeholders themselves.

Responsibility for persistence lies ultimately with the community requiring it and, although there are several variations in the practical possibilities for maintaining guarantees of persistence, having a governing stake over the long-term for a chosen scheme is essential

DiSSCo is an international undertaking spanning multiple institutions across many countries, each with their national interest. A governance mechanism must be sensitive to local/national issues as well as fitting with the nature of DiSSCo as a European research infrastructure. Considering also the desirability and potential for global adoption of a single scheme (below) flexibility for extension to worldwide governance is an important dimension.

The surest way to deliver trustworthiness and persistence, and commitment to those aims is through governance by stakeholders themselves with accountability to the wider community. This is in the interests of both the stakeholders and the wider community. Despite that delivery of scalability might be best achieved through outsourcing and subcontracting, responsibility for that and its governance still rests with the stakeholders.

Appropriate identifiers

Requirement: PIDs appropriate to the digital object type being persistently identified.

The digital object to be identified and the circumstance of the object's use dictate that PIDs should be appropriately chosen to reduce duplication, proliferation, and confusion of PID scheme types.

With differing circumstances of use and type specific metadata needs, DiSSCo has identified several categories of digital object, in addition to Digital Specimens that need to be identified (Table 1), with suggestions for possible PID schemes that could be adopted for each.

Potential for global adoption

Requirement: Extensible towards a single PID scheme that could be adopted globally.

A single worldwide PID scheme for identifying digital specimen data is highly desirable because collections-based science is global, even if collections are both managed and used locally. It will become increasingly difficult for scientists when different PID schemes are used in different countries or regions.

A Digital Specimen PID scheme that fits with current regional and institutional practices for physical specimens and local specimen records is desirable and possible. The success and acceptance of a PID scheme for Digital Specimens will depend not only on the persistence of the PID layer but (as with DOIs) in a large degree also on connection to stable local resources and landing pages, such as provided by CETAF Stable Identifiers and International GeoSample Numbers (IGS). Neither CETAF Stable Identifiers nor IGSNs represent Digital Specimens. They are conceived as identifiers for physical objects in institutions (Güntsch et al. 2017, Lehnert et al. 2019). This role will continue to be important after a PID scheme for Digital Specimen data is established. The scheme should be usable to make references to specimens, in literature or elsewhere that have not yet been digitized. This can be done by creating a PID for an 'empty' Digital Specimen object even before digitization has commenced.

Global extension of a chosen PID scheme does not alter the principal requirements around scalability, persistence and trust. However, consideration for how a regional (e.g., DiSSCo) level governance mechanism can interact with or evolve towards a worldwide governance model is an important factor to consider.

Available Handle-based PID schemes

The Handle System (Sun et al. 2003) operates on the basis of assigning responsibilities for administering portions of the entire Handle namespace. This namespace, of which every Handle is a member consists of two parts: a naming authority, represented by the Handle prefix, and a unique local name under a specific naming authority prefix, otherwise known as a Handle suffix. Naming authorities are organised hierarchically beginning with the DONA Foundation with, in principle no limits on delegation of responsibility. But in practice, delegation beyond two or three levels becomes unwieldy.

The known naming authorities and the different schemes they operate are each a variant of the Handle System. Eight of these are the available PID schemes considered in the present appraisal. Each is described in the following subsections, beginning with the most familiar scheme.

Holding a responsibility as a naming authority to administer one of the levels of the Handle System namespace and the relation between that responsibility and the naming authority at the next highest level is a specific business relationship that we name as the 'operational mode'. The possible operational modes we consider in the present appraisal for each of the available PID schemes are explained further below in the section on operationalizing a PID scheme.

Digital Object Identifier (DOI)

The Digital Object Identifiers (DOI) scheme is the well-known PID scheme widely used in academic publishing, research and other sectors. In publishing it has become instantly recognisable through its use of the ‘10dot’ prefix (as in this example: doi: 10.1000/182 or its

fuller DOI URL form https://doi.org/10.1000/182), its regular occurrence in article header information and bibliographies, and through marketing of the DOI brand. More recently, DOI is increasingly well known as an identifier for published datasets. Around 240 million DOIs have been registered to date though ten Registration Agencies (RA) that include Crossref, DataCite, EIDR and the Publications Office of the European Union. These RAs provide DOI registration, resolution and other services to their respective communities under different business models. The DOI PID scheme is governed and managed on behalf of its RA members by the DOI Foundation (IDF). Collectively the IDF and its RA members assume the long-term responsibility to maintain and sustain the DOI PID scheme for everyone.

International GeoSample Number (IGSN)

An International GeoSample Number (IGSN) is an alphanumeric code, obtained through one of the IGSN Allocating Agents, for uniquely identifying material samples from the natural environment together with their related sampling features. Like schemes such as the Deep Carbon Observatory Identifier (DCO-ID), it is a specific case of the Handle System five-digit prefix scheme (below) in which the alphanumeric code forms the suffix of a Handle with prefix 10273. With roots in the identification of geological samples, the IGSN system assigns identifiers mainly for non-biological samples such as rock, mineral and fossil specimens, dredges, cores, etc. Small numbers of biological and archaeological samples are known to have been registered as well.

At the time of writing (March 2021), the IGSN e.V. implementation organisation and its stakeholder community is concluding a strategic review project funded by the Sloan Foundation. A plan and roadmap - 'IGSN 2040' - will guide IGSN towards an mature future as the global PID scheme for material samples of all kinds, not only geo/earth samples. The work intends to re-design and mature the existing organization and technical architecture of the IGSN scheme towards a global technical and organizational infrastructure - the Internet of Samples (Davies et al. 2021).

European PID Consortium (ePIC)

The European PID Consortium (ePIC) provides PID services for the European and wider international research community under the top-level Handle prefix "21dot". The expectation is that the ePIC consortium will provide PID services to the European Open Science Cloud (EOSC) as a highly reliable, persistent and high performance service. Governed by a Memorandum of Understanding that aims to provide long term reliability, ePIC is a Europe-wide consortium of European information technology service centers for science. GWDG, one of the nine Handle system Multi-Primary Administrators (MPA) responsible for sustaining the Global Handle Registry Services is among the members and takes a leading role. The ePIC Consortium, especially by leadership of GWDG has been instrumental in devloping the persistent identifier policy and architecture for the EOSC (European Commission 2020, European Commission 2021a).

Five-digit prefix (CNRI)

CNRI Inc. as the founder of the Handle System and one of its Multi-Primary Administrators (MPA) with shared responsibility for the Global Handle Registry offers Handle prefixes and registration services for a small annual fee. An organisation requiring PID services enters into a Registry Service Agreement to act as a Local Handle Service Provider (LHSP) operating and maintaining software systems providing local handle services (LHS) for minting, registering and resolving PIDs under their registered prefix(es).

Historically, five-digit prefixes of the form “1nnnn” have been registered. Already mentioned examples including prefixes allocated for IGSN (10273) and the Deep Carbon Observatory (11121) but there are many others. Whilst these legacy prefixes can remain in use in the future, new registrations of five-digit prefixes are deprecated in favour of the second-level and three-segment prefix schemes.

Second-level prefix

The second-level (or two-segment) prefix scheme is the general version of the more specific DOI and ePIC administration schemes described above. In this scheme, authorised Registration Agencies act for one of the nine credentialled MPAs to administer a portion of the namespace for specific PID user communities under the relevant top-level prefix.

Of the MPAs currently authorised, three have global scope and relevance for DiSSCo, namely: The DOI Foundation, GWDG and CNRI. Respectively, these MPAs are presently responsible for the 10dot, 21dot and 20dot top-level prefixes. The remaining six MPAs all operate with territorial scope in countries outside Europe and are not considered further.

Three-segment prefix

Three-segment prefixes have replaced the deprecated five-digit prefix scheme (above). Three-segment prefixes allow a more granular level of prefix sub-division beyond second-level prefixes for administrative convenience to reflect organisational divisions and responsibilities. The current prefix allocation scheme operated by CNRI Inc. through their Handle.Net® Registry is a two-segment/three-segment scheme i.e., both are possible under a top-level '20dot' prefix.

The three-segment prefix 20.5000.1025 is presently assigned to DiSSCo for experimental purposes. Since 2019, DiSSCo has been using this for evaluation and testing of the Digital Specimen Architecture approach and its technology components through the DiSSCo Natural Science Identifier Registry (NSIDR) demonstrator. Such prefixes are an attractive option as they offer a low-cost path of least resistance to piloting DiSSCo operational services in the short term.

Two-digit top level prefix

At the highest level of namespace administration, DiSSCo could request a new distinctive two-digit top-level prefix specific for identifying Digital Specimen objects. This could sit alongside existing top-level prefixes such as those for DOI (10dot), general Handles (20dot) and European research (21dot) – to give just three examples.

Choosing this variant would require DiSSCo to adopt an operational mode of either allying with an existing MPA to obtain a delegated top-level prefix or to act as an MPA in DiSSCo's own right and be assigned a top-level prefix by the DONA Foundation, the non-profit body responsible for overseeing Handle System administration. The issues associated with adopting a new two-digit top-level prefix are complex – organisationally, financially and politically.

National-level services

There is increasing evidence of emerging national-level support and services in several countries for persistent identification in the higher education and research sector. In part, this is due to the 'allocating agent' member model applied by RAs such as Crossref and DataCite. Much more so, it is due to recognition at the national level of the importance that PIDs play in connecting up the different parts of the research (information) ecosystem and as part of countries' commitments to making publicly funded science openly accessible.

France, for example, in its National Plan for Open Science (MESRI. 2018) has declared national commitment to making open science a normal part of everyday practice for researchers; and to helping to define and regulate the building blocks of the open science ecosystem, such as Crossref and DataCite for DOIs and ORCID for researcher identifiers. This kind of declaration is typical in many countries.

In Germany, the ZBMED Information Centre for Life Sciences is a member of the DataCite consortium, taking the role of a DOI broker or allocating agent for German research institutions in life and medical sciences. The German National Library of Science and Technology (TIB) provides the underlying technical infrastructure and offers a similar service with focus areas in engineering and technology.

In the Netherlands, the cultural heritage and the library communities are active adopters and users of different types of persistent identifiers. The most common PID types in the Netherlands are DataCite DOI (operated through Delft University of Technology), the ePIC Handle system (through SURFsara) and URN:NBN (though the National Library and DANS). The National Coordination Point for Research Data Management organizes various workshops and training focusing on different aspects of research data management.

In the UK, Jisc is working to select and promote a range of unique identifiers in collaboration with relevant partner organisations and funders of research, who may (for example) consider mandating the use of such PIDs as a condition of research grants.

In many of these cases, the effort is mainly directed towards identifying datasets, people, organisations, research grants and research outputs. Only the UK's HeritagePIDs project, as far as we are aware is investigating approaches to PIDs for institutions across the UK heritage sector as the sector considers how to work towards an open national collection.

As a Europe-wide research infrastructure with a new requirement to persistently identify Digital Specimens, DiSSCo will be best positioned by selecting and promoting identifiers with relevant partner organisations on the European rather than national level. For other kinds of identifiers (DOI, ORCID Id, GRID/ROR) it makes little difference whether these are provided/used on European or national level. This is the matter of operationalizing a PID scheme, explored further in the following section.

Operationalizing a PID scheme

Operational modes for Handle-based PID schemes are based on the kind of responsibilities associated with administering part of one of the levels of Handle System namespace (Sun et al. 2003).

A root or top-level name segment ('10.' in the case of DOIs, for example) is administered by a Multi-Primary Administrator (MPA). Presently there are nine MPAs constituted. Collectively, they operate the Global Handle Registry (GHR). Being an MPA carries a substantial commitment and long-term obligation towards the GHR with responsibility for ensuring (collectively) the continued governance, operability, sustainability and persistence of the entire Handle System. Among the PID schemes described, the DOI scheme, the ePIC scheme, the 5-digit prefix (CNRI) scheme and the two-digit top level prefix scheme are all examples of primary PID schemes, each administered by a responsible MPA.

At the next level down, segments of the namespace are administered by Registration Agencies (RA), typically on behalf of identifiable communities. Crossref, for example, manages a segment of namespace below '10dot' in the DOI scheme on behalf of the journal and scholarly publishing sector. DataCite manages a similar segment for dataset registrations on behalf of the global research community. However, neither Crossref nor DataCite allocates DOIs directly. They delegate that further to their members acting as agents.

Among the described PID schemes the generic second-level prefix scheme is an example having similar arrangements to the DOI and ePIC schemes i.e., an MPA administers the top-level name segment with one or more RAs administering second-level name segments. In the generic three-segment prefix scheme this is taken to one further level of administration with individual institutions managing their specific third-level name segment (prefix) allocation.

The IGSN scheme for sampling communities is a variation on the above general principles in which a Registration Agency (IGSN e.V.) has administration rights for a top-level name segment delegated to it but with the GHR obligations being retained by the MPA responsible for the name segment. Note, this is a specific example of the legacy five-digit

prefix (CNRI) scheme. In the IGSN scheme a further sub-division of administration is devolved to multiple IGSN Allocating Agents (AA).

Several possible operational modes for DiSSCo can be distilled from these different scheme options. These fall into two main categories, being: i) administration of a root or top-level name segment; and ii) administration of a lower-level name segment under a top-level name segment. Suppl. material 3 presents a short comparison of the main roles and responsibilities within these two categories. The choice for DiSSCo depends in part on the value placed strategically on identifiers for digital specimens on the Internet, as explained above; as well as on the practicalities and socio-political realities of implementation in a community that is still learning to be digitally progressive.

In both main categories there is a further choice about how much to operationalise inhouse versus how much to outsource (subcontract) to another organisation to carry out on behalf of DiSSCo. Note though, that choosing an outsourcing option does not outsource the accountability for correct administration. That remains with DiSSCo. Also, several technical facets, such as maintaining metadata schemas and accuracy of Handle records nevertheless also remains with DiSSCo.

A further consideration is that the principal functional roles of a PID scheme (i.e., PID registration, PID maintenance and PID resolution) must each be assigned to an entity or entities within the foreseen DiSSCo organisational structure. Should, for example these roles be performed centrally by DiSSCo on behalf of all participating institutions? Should some or all of the roles be devolved, and if so, which roles and to whom? Is a mixed model anticipated?

Administration of a top-level name segment

Within the overall category of choosing to administer a root or top-level name segment there are two options: to ally with an existing MPA or to become and act as an MPA.

There are presently nine MPAs authorised by the DONA Foundation, collectively sharing the responsibility to operate the top-level Global Handle Registry (GHR). Six of these are territory oriented and can be discounted for alliance purposes whilst the remaining three (IDF, GWDG, CNRI - see Table 2) have non-territorial global scope and are worth considering.

Ally with an existing MPA

DiSSCo could negotiate an alliance with one of the existing non-territorial global MPAs; either to:

1. Make use of their existing top-level prefix:
    - IDF/10dot
    - CNRI/20dot
    - GWDG/21dot; or

2. Sub-contract the MPA to manage a new top-level PID prefix on DiSSCo's behalf.

There are precedents for sub-contracting already. The DOI Foundation is an MPA paying full dues to DONA Foundation as a constituent contributor to the Global Handle Registry. However, all DONA-level administration is sub-contracted to CNRI as well as much of the lower level administration for DOI. Subcontracting is also a model within the IDF at the RA level. The EU Publication Office (OP) is a full-fledged DOI RA but all DOI operations are subcontracted to mEDRA, the Italian DOI RA. In each case, OP and IDF, the organization needed to be a first-class citizen for political reasons but didn't want to build up the internal structure needed to operate as a first-class citizen.

The base annual cost of these two alliance options differs depending on whether DiSSCo uses an existing top-level prefix or a new prefix. In the latter case, the base cost is most likely the same as being an MPA – annually, CHF 75,000. In the former case, it's probably negotiable.

Act as an MPA

Acting as its own MPA allows a community to establish their own specific Handle-based PID scheme with the benefit of being able to customise and control the scheme exactly as the community likes. This makes it easier to meet specific community requirements without compromise.

Becoming one of the exclusive, small number of MPAs at the global level carries obligations, however. It requires the MPA organisation to commit to contribute (alongside the other MPAs) to sustain the Global Handle Registry (GHR) for the benefit of all countries, sectors and individual participants. This is expensive and may appear beyond the immediate interests and needs of DiSSCo. The base annual cost of being an MPA is CHF 75,000 annually payable to the DONA Foundation to keep that Foundation and the GHR operational. It is unlikely DiSSCo would take on such a responsibility on its own.

Nevertheless, when viewed in terms of DiSSCo as the largest ever formal agreement between natural science collection facilities, the many thousands of collection-holding institutions globally, and the overall general importance of guaranteeing very-long term (100 years) persistence of digital references to all kinds of heritage objects, this can take on a different complexion. DiSSCo might consider doing so in collaboration with other natural science collection digitization initiatives around the world or in conjunction with other partners in the wider heritage collections and infrastructures sector generally.

Administration of a second-level name segment

Within the overall category of choosing to administer a lower-level name segment under a top-level name segment there are three options: to use the services of an existing RA, to ally with an existing RA or to become/establish a new RA.

A Registration Agency provides services to those wanting to register PIDs. These include allocation of prefixes within the range administered by the RA, registration of PIDs and capture/maintenance of associated metadata.

Use the services of an existing RA

DiSSCo could use the offered services of an existing RA, such as CrossRef or DataCite in a customer/supplier relationship. There are several existing RAs to choose from. Like domain name registrars for the Internet, RAs offer a range of services targeted towards their potential customer base for the registration and management of PIDs. They can compete with one another. Each RA manages their own business operation independently from other RAs. They can also collaborate with one another, both on mutually beneficial service development and as a collective in support of the goals of their parent MPA.

Ally with an existing RA

In a more strategic alliance, DiSSCo could work with an existing RA to influence and develop the RA's current services (especially the metadata schemas), operations and the business model to more specifically meet DiSSCo needs. A concern for an existing RA might be that the volume of business deriving from the collections' PID needs could potentially be disruptive to the current mode of operation/collaboration. A concern for DiSSCo is that a collaboration might involve compromises with the entire communities served by an existing RA.

Become/establish a new RA

Becoming an RA (i.e., establishing a new RA) under one of the existing MPAs is a flexible solution that allows PID services to be tailored to specific communities to meet specific needs. RAs serve their own communities and support them with bespoke metadata schemas, local language customisations and a business model specifically designed around the supported community.

A new RA could exploit an existing brand of its parent MPA (such as the DOI or ePIC brands) thus benefitting from the reputation of that brand already established. Alternatively, a new RA can establish its own distinctive brand, such as MovieLabs chose to do in the entertainment sector with the Entertainment Identifier (EIDR) brand and as BSi.Identify will do when it launches as an RA in 2021.

Material and methods

By combining the possible operational modes with the different available PID schemes explained above, Table 3 illustrates the many scenarios of operational PID scheme from which DiSSCo can choose. However, not all are possible.

Combinations shown as 'not possible' (A2, B1, B2, B5, C2, D2, E3, E4, F2, G1, G2) are not operationally and/or organisationally achievable. Two scenarios (D1, D5) are deprecated because the offering from the service provider has changed and three at the national level (H3, H4, H5) are not desirable to pursue as they will lead to fragmentation along national lines. This leaves the remaining twenty-two 'possible' scenarios that have been subjected to analysis against the DiSSCo requirements in a two-step approach.

In a first step aiming to reduce the number of alternatives to a sensible and practical subset, a coarse three-level strength/weakness scoring has been applied to each scenario against each of the principal DiSSCo requirements outlined earlier i.e., scalability, community trust, persistence, governance, appropriateness of the scheme and suitability for future global adoption.

The eight strongest scenarios based on these ranked scores were selected to proceed to the second step of a more detailed dimensional benefits assessment. These assessments were made under several headings:

- **Option summary**: An explanation of the option and how it operates.
- **Outcomes**: The outcomes likely for DiSSCo and the community from adopting the option.
- **Impact**: Impact expected to be achieved by adopting the option.
- **Implications**: Implications for DiSSCo (collectively) arising from adopting the option; also including obligations DiSSCo would have to commit to over the long-term (20+ years).
- **Pros**: A list of positive aspects of the option.
- **Cons**: A list of negative aspects of the option.
- **Costs**: A non-exhaustive indicatication of the costs of the option, particularly their type and scale (see note below).
- **Dimensional assessment**: Ten dimensions (characteristics) analysing the ability of the option to meet the key requirements. Each dimension is assessed on a three-point scale: i) able (appears to be able to fully meets the requirement); ii) partial (appears partially able to meet the requirement); and iii) unable (appears unable to meet the requirement). The ten dimensions are:
    - Billions of identifiers (several tens – 300 billion);
    - Flexibility for machine-assisted services;
    - Consistency (to continuously sustain services over long-term);
    - Quality (conformance to stakeholder requirements);
    - Excellence (towards convergence of curation and publishing practices);
    - Scope for branding own scheme;
    - Persistence (100 years resolution);
    - Opportunity for stake in long-term governance;
    - Flexibility to accommodate specific metadata in PID records; and,
    - Suitability for expansion to a global scheme.
- **Overall assessment**: Considering all factors, an overall assessment of the option is also indicated as able, partial or unable. An additional remark derived from the pros and cons of the option qualifies the assessment.

Note on costs: In providing indications of costs we have tried to identify the major sources and components of costs without being a complete costing. The idea is to provide a scalar comparison along the lines of 'low, medium, high' or 'affordable/unaffordable'.

Results

A three-point scoring scale has been used in a first step to assign strength/weakness scores to the ability of each PID scheme scenario to support each of the main DiSSCo requirements (Table 4). The scores assigned (where 3 = strongly meets the requirement, 1 = weakly meets the requirement and 2 = inbetween, neither strong nor weak) are summed for each scenario, giving an overall score for each in the range 18 (strongest) – 6 (weakest).

These scores show that, operationally speaking allying with an existing MPA is preferential to becoming a new MPA; and that becoming or establishing a RA is more likely to lead to satisfactorily meeting DiSSCo's needs than using the services of or or allying with an existing RA. None of the six options to use the services of an existing RA (options identified with digit 3 in Table 4) has been selected for further evaluation because the consequence would be that DiSSCo would have to fit in with existing services and metadata schemas that most likely would lead to inflexibilities and/or inability to fulfil the main requirements for FAIR, scalability and persistence and governance/trust. Working within one of the available PID schemes from DOI Foundation, ePIC Consortium or CNRI seems to have advantages over the other schemes (IGSN, five-digit prefix, two-digit top level prefix, second level prefix, three-segment prefix, national level services.

The overall scores from Table 4 carried over into Table 5, identifying the strongest scenarios (in ***bold italic***) for the second step of further evaluation, which also includes evaluation of a 'do nothing' option. The details of the assessment of these strongest scenarios are given more fully in Suppl. material 4.

The scenarios to ally with the DOI Foundation (option A1) and to become an RA member of the DOI Foundation (option A5) were assessed as able to fully meet DiSSCo requirements, with DiSSCo especially being able to benefit from the familiarity of an established DOI brand. The scenario to become an RA in association with the ePIC Consortium (option C5) was assessed as able to meet the DiSSCo requirements but with a less well established and familiar brand than DOI. The scenario to ally with IGSN (option B4) was assessed as being partially able to meet DiSSCo requirements but fully able only after substantial cooperative work between DiSSCo and IGSN stakeholders.*[6] The scenarios to become an MPA or to ally with an existing MPA for a new top-level prefix (options E2/E1) were assessed together and found to be able to meet DiSSCo requirements but presently beyond the capability reach of the DiSSCo community at the present time; and with a potentially high risk of failure.*[7] Allying with another MPA for a second-level prefix and become an RA under that MPA (options F1/F5) were assessed together and found to able to meet DiSSCo requirements but with a medium risk of failure due to the greenfield nature of the scenario and strong community memories in relation to the failed life sciences

identifier (LSID) scheme. Becoming a Registration Agency for a three-segment prefix (option G5) was assessed as able to meet DiSSCo requirements but most likely with limited flexibility for further devolving namespace management whilst minimising verbosity of Handle names and maintaining opacity of suffixes. The 'do nothing' option, which represents no change to the present situation does not meet DiSSCo requirements at all.

The scoring of each scenario is carried forward in illustrative form into Fig. 2.

Discussion

The scenarios of establising an RA under the DOI, ePIC, Two-digit top level prefix and Second-level prefix schemes come out as the strongest options from the assessment. The option of DiSSCo establishing itself as an authorised Multi-Primary Administrator can be discounted because of the implied long-term obligations. The option to ally with an MPA such as the DOI Foundation in some manner is an opportunity not ruled out.

Aligning a Registration Agency (RA) to a brand

The RA scenarios (A5, C5, E5, F5) are attractive on several levels. They offer the promise of being able to develop and acquire the trust of the target community through recognition and confidence in a relevant RA market brand. This can be branding created by DiSSCo itself along the lines that RAs like EIDR and BSi.Identify have achieved or branding in combination with the brand reputation of the top-level prefix administrator (DOI Foundation, ePIC consortium, Handle.net). To what extent is it helpful to leverage an existing brand versus developing an own brand such as 'Natural Science Identifier (NSId), for example? DOI is a very strong trademarked brand. ‘Driven by DOI’ or ‘DOI powered’, for example can be an enticing strapline that conveys confidence that DiSSCo is not doing much different or out of step with other, well established uses of the DOI brand. The ePIC brand is likely to grow in the context of the EOSC. ePIC services are not only European but global in scope and eventually will be tailored specifically towards the research (data and infrastructures) communities. Handle.net® is widely used but is better known in technical circles than non-technical ones. Brand fit has to be considered and there is no doubt there is a good fit with DOI even if DiSSCo takes its own branding route too. “Natural Science Identifiers, driven by DOI” has a reassuring confidence about it, suggesting that DiSSCo is buying into an established mechanism whilst offering the custom characteristics needed by the sector.*[8] Operational compatibility with the journal and data publishing businesses, as well as with EOSC is assured; as are persistence and sustainability through the IDF with the support of its RA members, CNRI and the industry sectors that rely on DOIs now.

On the other hand, DiSSCo should remain aware of the increasing importance of DOIs in and to the commercial sector. There is a potential risk to IDF and its non-profit RA members of increasingly coming under commercial influence that can grow stronger in the future. That might be an argument for a different scenario to avoid that the publicly-funded heritage collections experience pressures that are disadvantageous. This could push DiSSCo more in the direction of ePIC or Handle.net if that is thought to be a problem. Or it

could dictate that as well as acting as an RA, DiSSCo should aim to ensure a strong alliance and influence in IDF.

Handle.net is a weaker brand model, sustained by CNRI and its business. RAS (Reliability, Availability, Serviceability) policies and plans are unclear although CNRI is known to be working on this. CNRI Inc., is a privately held, USA-based not-for-profit organisation established to "*undertake, foster, and promote research in the public interest.*" As the founders of the Handle system and operator of both local Handle registry services (Handle.net®) and a significant part ofthe Global Handle Registry as well, CNRI has a large stake in play. CNRI also offers subcontract technical and operational services to some of the other players, including operating the DOI infrastructure on behalf of IDF. CNRI's 'HDL enabled' logo and strapline are not as well known as the DOI brand but technical capability is strong and, subjectively, resilience seems reasonable. It is important to note that the DOI system uses Handle.net as a necessary but not the only component for the DOI system.*[9] CNRI thus is sustained through the DOI ecosystem as well.

With its strapline of 'persistent identifiers for eResearch', the European PID Consortium (ePIC) is backed by some of the strongest research sector players in Europe, including CSC (Finland), KTH (Germany) and SURFSara (Netherlands), as well as the Swiss National Computer Centre, the German Climate Research Centre and the Greek Research and Technology Network. ePIC will probably gain more members and strength as EOSC becomes more established. Ultimately, it will become a reliable service provider with a specific research community focus.

Could DiSSCo as an entity live with working under an existing top-level prefix (10dot or 21dot)? The current consensus is that DiSSCo could do so, especially if DiSSCo were to establish its own distinct RA. That model has clearly been successful for Movielabs with EIDR, which is completely differentiated from anything DataCite and Crossref do. There's no confusion because EIDR is the brand and the users don't care that it's 10dot or that 10dot is also in use in other sectors. NSId as a 10dot brand isn't unappealing, perhaps with a distinctive second-level prefix such as '10.22/' or '10.30/' for example. This is a solution to take advantage of the DOI ecosystem and yet still allows DiSSCo to create its own unique service space. And that might be a deciding factor for the IDF in preference to GWDG/ePIC or Handle.net, whose top-level prefixes 21dot and 20dot and associated service spaces are not so well developed as yet.

The value of community specific identifiers

An important concern for the present options appraisal is to understand the opportunities presented by naming and adopting a community specific identifier like 'Natural Science Identifier' (NSId), the value to assign to these and the implications arising in terms of commitments to ensuring that digital specimens on the Internet remain openly findable, accessible, interoperable and reusable over many decades (100 years' time). How DiSSCo and its counterparts globally opts to play in the Handle System is ultimately a strategic decision about the value and branding of digital identifiers for digital specimens on the

Internet. How does the European and worldwide natural science collections community want to be perceived in an increasingly digital future?

Community specific identifiers and digital identifier systems more generally do play an essential role by offering a focus on uniqueness and persistent identification of digital objects that leads to stable, authoritative packages of scientific information and trust. On the other hand they are the face of a resolution mechanism for instant and online access that makes (with software assistance) navigation and interactive use of assets easy. Assigning identifiers to digital artefacts and associating relevant metadata with such identifiers has been shown to substantially increase findability and use of artefacts (Khedmatgozar and Alipour-Hafezi 2017). This is in keeping with the tenets of the widely promoted and modern movement that aims to make open science a normal part of everyday practice for researchers and to make scientific research results open to all – researchers, companies, citizens.

Transversely, a strong role is placed on digital identifiers as enablers of machine-actionable support to humans in an age of Internet diffusion, workflow automation, data explosion and machine learning/AI. This has long been the case in software systems of all kinds using unique identifiers internally – much of the time hidden from direct view. This is a practice that dates to the 1980s in origin with the invention of graphical workstations. Only now do the descendants of such internal identifiers assume a much more prominent and public role thanks to the Internet, digital transformation and the need to manage – as opposed to just see and move – data on the Internet.

Digital identifiers are enablers for social and professional change in the way in which collection-holding institutions both manage and share their holdings and expedite the related data into the hands of users. Thus, while perhaps not being able to assign a quantitative value to identifiers, we can say qualitatively they already have a high value now. Coupled with transforming physical specimens' data to digital specimens on the Internet, they have far more potential in coming years.

The question for DiSSCo and global counterparts is: How important is it to have a clear brand association between the objects identified and the identifiers themselves? What does that look like? What does it mean in terms of trust and persistence (and ultimately, governance) for a chosen PID scheme and its mode of operation?

Assigning value in terms of trust and persistence has a consequence both for the choice of PID scheme and for the choice of an operational mode within the selected PID scheme. The two choices are not completely independent of one another. How does such value need to translate to a clear brand association between the objects identified and the identifiers themselves? And, discussed above how important is brand value?

That must be thought about in the context of the future value chains founded on natural science collections, especially those in recently described notions of extended specimens (Webster 2017), next generation collections (Schindel and Cook 2018) and the role of

collections in ensuring critical research and education in the current century (National Academies of Sciences, Engineering, and Medicine 2020).

'Open Digital Extended Specimens on the Internet' (openDS) enhance the value chains founded in natural science collections. These value chains extend from initial gathering and organisation of specimens, through conduct and commercialization of specific science based on specimens, to sharing ensuing economic and social benefits in a fair and equitable way. Products of digital value chains can provide the evidence for regulatory processes in health, food, security, sustainability and environmental change, and new educational uses. Future software applications can work with and on Digital Specimen objects to provide more sophisticated computer-assistance to both the present day known work tasks and to unimaginable future works of collection specialists, scientists and others working daily with specimens. Natural science collections and the science and other values that flow from them are part of the common good that must be shared with everyone. Open digital specimens on the Internet, persistently identified in a manner that makes them instantly recognisable to end-users – researchers, companies, citizens – are a specific category of raw materials for new knowledge generation (research) and acquisition (teaching and learning).

The preferred PID scheme

In consideration of the foregoing, the strongest option across the studied major dimensions of the available Handle System PID schemes and operational modes is for DiSSCo to use DOIs to identify Digital Specimens. The case for choosing DOI comes out slightly more strongly than choosing ePIC for reasons related to the substantial achievements, operational experience and reputation of DOI/ IDF to date. Operating under another Handle-system prefix than those used by IDF and ePIC is the substantially weakest option because of the difficulties associated with introducing an identifier that is not perceived to be a DOI. The term 'DOI' is trademarked by the IDF and thus not available for describing other identifiers.

The practical and sensible avenue to explore further are the options to establish and become an RA member of the DOI Foundation (option A5) and to enter a strategic alliance at the level of the DOI Foundation (option A1). These options are likely most effective when actioned in combination.

In a global context

Although DiSSCo is a European endeavour for digital unification of European natural science assets, the present options appraisal should also consider relevant developments from elsewhere, especially outside of Europe noting again that one of the main requirements for DiSSCo's preferred PID scheme is potential for global adoption. Here follows a dissection of relevant facts.

GBIF has been using DOIs for identifying occurrence datasets and queries for many years. GBIF is investigating what is needed to persistently identify each individual occurrence

record in a resolvable manner. The Darwin Core term 'occurrenceID' is already an identifier of each individual occurrence but in many cases this is not fully and directly resolvable back to the original data about that occurrence. Identifying an occurrence with a DOI or IGSN could solve that problem. GBIF has been investigating both schemes.

International Geo Sample Numbers (IGSN) have gained traction in the geoscience community. The iSamples Research Coordination Network (RCN), 2014 – 2019 was successful in laying the groundwork for the expansion of the IGSN scheme to other domains where material samples are of importance, such as natural and environmental sciences, material sciences, agriculture, physical anthropology, archaeology and biomedicine. In June 2020 a follow-on iSamples project received funding to build infrastructure to begin this expansion and engage new communities over the coming three years.

The Arctos collection management system (CMS), which curates and serves data from 180 separate collections can allocate DOIs to specimens recorded in that database.*[10]

The Specify consortium may decide to add functionality to the Specify collection management system (CMS) to begin adding DOI prefixes or IGSN prefixes in front the GUIDs that Specify CMS already generates for each specimen record. That would probably be easy for them directly (or with a collaborating third-party) to implement and deploy e.g., with a 'PID registration plugin' such as the IGSN compatible plug-in ("iSamples-in-a-box") that will be developed in the previously mentioned iSamples project. Apart from installing the plugin, a collection manager would only need to sign up with a registration agency such as DataCite to begin registering DOIs.*[11] The cost of that would be perhaps $5-10k per year to an institution. Specify is one of the most widely used CMS software solutions, especially in North America. Presently there are more than 275 installations of the software, managing 450 collections across 38 countries.

The CMS landscape is fragmented. In a survey by DiSSCo in 2017 (Casino et al. 2019) it was found that many institutions use more than one collection management solution. 117 systems were reported by 89 institutions. Often, in-house custom solutions (36%) or solutions based on MS Access, MS Excel, MS Word or Filemaker (21%) are used. The top 3 CMS found were Specify (7%), Adlib (3.5%) and JACQ (3.5%). In a recent survey (Spring 2020) of 200+ respondents presented at the 2020 virtual conference organised by the Society for the Preservation of Natural History Collections (SPNHC) (SPNHC & ICOM NATHIST 2020), the top three CMS in use were found to be: Specify (23%), Axiell EMu (13%) and Arctos (12%), followed by 'Microsoft Access/Excel & Filemaker' (25%). Custom and minority solutions account for the remainder. That survey was mostly an anonymous straw poll so the responses were technically international; but for the fifty or so people contacted after completing the survey, most seemed to be from North American institutions. Thus, it's likely there's North American bias in the result. Nevertheless, taken all together the results illustration that multiple different CMS softwares are in use and the domain is ripe for churn for new capabilities.

Despite that only one DiSSCo member institution makes use of Axiell EMu and that Arctos is not used by any institution in DiSSCo, what Specify, Arctos and EMu decide to implement in terms of PIDs for specimens could have an important bearing on directions in general and thus for DiSSCo and global counterparts. If DOI and/or IGSN enhancement in CMSs were to go ahead, that would be quite impactful. The main iDigBio players use mainly Specify so this would quite suit them and probably VertNET as well.

Lastly, it is important to recall in this context the similarities and differences of DiSSCo's approach for the digital representation of physical specimens (Hardisty 2019, Hardisty et al. 2020) with the Extended Specimen Network (ESN) concept (BCoN 2019, Lendemer et al. 2019). In the ESN concept presently, digital representations originate and root in the CMS record itself, adding external resources to that. Both Arctos and Specify CMS are at least partially capable of making those linkages to external sources and that capability can grow. But this leads to more complexity within and across such CMSs. Extensions and harmonisations are in the hands of the different CMS vendors and development consortia. Enhancing Specify records with a Handle prefix (as already done by Arctos) is already then persistently and uniquely identifying extended specimens. A PID per specimen record could be registered, made up of a prefix and the GUID that Specify already automatically generates for each record. DiSSCo's Digital Specimen concept is different here, with each DS as a digital entity distinct and separate from (external to) to a specific CMS record. This involves a separate/new PID to identify the DS alongside, for example the CETAF Stable Identifier identifying the physical specimen and its corresponding collection management database record. The DS and the physical specimen remain coupled through the information associated with the PID of the DS. CMSs like Specify and Arctos can easily be adapted to point in the other direction to corresponding DSs as well. Discussions to achieve technical convergence, as we explained earlier are in progress*[1] with the expectation that this will be achieved. 'Digital Extended Specimens' will be distinct, identified digital entities that represent specimens on the Internet, extending the information about them normally held in institutional collection managment systems. They will be separately processable.

Steps to implementation

In the simplest sense, offering a PID service means providing, on the one hand a resolver that provides a table lookup with redirection to the asset of interest, and on the other hand a mechanism for capturing information into that table to make the resolution possible. However, over the years, the global scientific community, data and service providers realised providing such service is not such a simple endeavor. Well-founded PID initiatives have focused on building robust, trustworthy, reliable and sustainable service that can cater to the global research community.

DiSSCo needs to *warm up its engines* to begin delivering pilot and pre-production level services whilst at the same time consulting with global counterparts to develop the necessary governance, finance, operations and architecture that can ultimately lead to a worldwide PID scheme for Digital Extended Specimens.

In the time prior to a DiSSCo RA becoming operational (c. 18 months from decision), DiSSCo will join IDF as a member and work to pilot the simplest PID service whilst developing the four GOFA areas (governance, operations, finance, architecture) and processes as prelude to an RA member application procedure to the IDF. This involves developing a service model that fits the DiSSCo service portfolio guidelines with key performance indicators to measure success. Preliminary elements of some of this necessary work are mentioned below.

Governance, finance, operations and architecture

Illustrated in Fig. 3, four interconnected 'GOFA' areas*[12] together deliver the long-term sustainability needed by the DiSSCo PID scheme:

- Governance: How an RA operates, how it controls and manages decisions on PID services. How is membership and stakeholders defined and managed. Other scope of governance: roles and responsibilities, liabilities, standards, and policies.
- Operations: The processes, tools and expertise needed to run a PID service.
- Financing: The resource needed for building, maintaining the architecture and operating the service in a reliable and trustworthy manner.
- Architecture: Refers to the technical design specifications about the PID service.

These GOFA quadrants are interrelated. Choices in one quadrant affect the others, as illustrated by several examples: A membership model that relies on membership fees might need a cost model that requires other financial resources (subsidy, in-kind) for maintaining long-term service commitments. A design decision to create a robust, high-availability server cluster requires significant investment in hardware, software and expertise and thus has a direct impact on operational costs. Specific PID policies can have an impact on operational and design procedures (e.g., access control, API design etc.).

Extending the ideas for operationalizing a PID scheme, we look at requirements, roles and responsibilities for becoming an RA. This helps us understand operationalizing a PID scheme as part of the future DiSSCo service portfolio and within the context of the EOSC landscape. We can think of the list following as the minimum/essential elements that are required for an RA to deliver a PID service:

- Governance:
    1. Provide governance and membership structures.
    2. Provide service terms and conditions (taking into account an SLA with the operational contractor if outsourcing is used).
    3. Provide PID policy (covering, for instance how the RA guarantees the persistence of the PID, how to handle PID updates, how to handle long-term preservation of the metadata associated with identified objects).
- Finance:
    1. Business case and cost model in place.
    2. Long term financial support outlook.
    3. Cross-cutting risk analysis and mitigation plan.

- Operational: (can be outsourced via a contract and service level agreement, if desirable)
    1. Process and procedure descriptions; operational handbook.
    2. Technical infrastructure: Provide reliable resolver (Local Handle Service - capability that a persistent identifier can be resolved to an object such as file or webpage).
    3. Technical infrastructure: including mirroring, redundancy, backup, archival capabilities.
    4. Technical infrastructure: Workflows and supporting software for metadata collection, PID assignment and registration.
    5. Provide human and machine-actionable (web services and API) interfaces.
    6. Provide domain specific and customisable Metadata profile and mechanism for profile creation, maintenance, update.
    7. Value added services: citation tracker, reporting, services building on PID graph, etc
    8. Provide user support (community engagement, stakeholder management, training and education).
- Architecture (design):
    1. Guarantee that a persistent identifier is unambiguously assigned to a resource within the system (this will be achieved by adopting the Handle system).
    2. Uncoupling hosting from identifier management (this will be achieved by adopting the Handle system).

Indicative costs

There are several ways, practically that an RA can be set up, with the not-for-profit basis being a common choice. For comparison purposes*[13]:

- **Crossref:** Established for twenty years now, the not-for-profit revenue from Crossref's membership fees and service charges has risen steadily from €3.8 million in 2010 to €7.7 million in 2019, growing by around €430k per annum. Crossref has more than 11,000 members.
- **DataCite:** Turnover from fees (again, not-for-profit but a smaller RA than Crossref) was €361k in 2017, growing by roughly €245k annually to €848k in 2019. DataCite was established in 2009. Presently, DataCite has around 200 members.
- **EIDR:** Founded in 2010, the Entertainment ID Registry Association (EIDR) is turning over €995k (average) across the years 2015 - 2018 (latest for which figures are available). Presently, EIDR has 70+ members.

Taking DataCite and EIDR as being more comparable for DiSSCo than Crossref (which serves a very broad base of different customer types) these numbers suggest that a DiSSCo RA, which ultimately might serve several hundred to two thousand members might reasonably be configured as a €1 - 1.5 million per annum not-for-profit business over the medium term.

A mixed model of funding will encourage membership. The model will need to begin simply and evolve, taking account both of the phasing of the DiSSCo programme and expansion of the PID scheme to the global level. Start-up and early operating costs are likely to come from the DiSSCo operational budget or from contributions of a small number of founding institutions. New members joining later may be asked to pay membership fees. As the number of members grows, fees should decline for everyone since the core service will mostly be fixed cost from the beginning. Additional, value-added services launched later might be subject to separate pricing models to cover their operating costs by those using them directly.

The cost to DiSSCo of the PID scheme must evolve in line with for the planned sequence of activities to take DiSSCo through its early pilot, implementation and full operational phases, anticipated as follows:

- **Early pilot phase (2022 - 2024)**: PID systems, processes and procedures will be put into place and operated on a trial basis for specimen indexing and for arranging loans and visits. PIDs minted during this phase will be guaranteed resolvable over the long-term.
- **Implementation phase (2024 – 2026)**: As confidence builds in trial results, procedures and systems DiSSCo will commence wider service deployment and controlled scaling to meet growing user demand.
- **Operational phase (2026 onwards)**: Final PID systems processes and procedures will be operated at full intended scale for the operational lifetime of the DiSSCo infrastructure.

In each phase, component costs comprise at least the following:

- Fees due from DiSSCo to another organisation for membership/administration of chosen PID scheme;
- PID minting and resolution fees (as appropriate);
- Service/system purchase, installation, running and maintenance costs (hardware and software);
- Personnel costs (system, process and procedure administration);

Assumptions include the need to have dual redundant systems or mirrored load-balanced systems, and multiple trained personnel (at least four capable of operating the system, with succession and risk planning).

Measuring success

Planning and investment must focus on creating a PID services model that is fully aligned with FAIR and EOSC recommendations (European Commission 2020). Such a model is illustrated in Fig. 4. Along with the technical infrastructures (for example, local handle servers and mirrors, Digital Object repositories) detailed service specification, service management plan, and community engagement initiatives (training, workshops,

hackathons, etc.) will be essential components for successful deployment and operation of the PID services.

By default, the Digital Specimen / PID services model combination is FAIR aligned through DiSSCo's choice of Digital Object Architecture as the technical basis (Lannom et al. 2020). The Data Management Plan for DiSSCo (Hardisty 2019) further assures FAIR compliance by designating 'FAIRness' as a protected characteristic of the DiSSCo infrastructure. This is a significant aim and advantage already partially achieved that can be further assisted by defining appropriate key performance indicators in relation to the benefits outlined earlier in the present article. Some suggestions are given in Table 6.

Further KPI development work is foreseen to ensure that PID services comprehensively support important data stewardship matters arising from requirements related to access and benefit sharing, data sovereignty, and data/knowledge rights of indigenous peoples.

Conclusions

DiSSCo has examined the PID needs to support a FAIR Digital Objects Architecture approach as the main path to implementing the data architecture of the new DiSSCo data infrastructure. Since 2018, DiSSCo experts have been identifying the requirements to be met by a PID scheme to support the concept of the Digital Specimen. The Technical Team of DiSSCo (the present authors) has produced the present options appraisal document evaluating, from a technical and social point of view, the options available in a framework to set up the conditions for choosing a PID scheme for natural science collections. By analyzing and assessing the relevant global PID landscape and several global and national PID service organisations, we identified key elements that are needed to choose a PID scheme for the DISSCo community.

The recommendation is to adopt a DOI-driven approach ('driven-by DOI'®) for the persistent identification of Digital Specimens. This approach leverages the achievements and acceptance of the widely recognised DOI trademarked brand for Digital Object Identifiers and is fully compliant with the FAIR principles. It builds on current science-policy and technological recommendations for the further development of the European Open Science Cloud (EOSC). As well as being aligned with current practices across the community of natural science collections, it is aligned to practices throughout the wider research and scholarly community and would be suitable for adoption more widely across the heritage collections sector.

A driven-by DOI approach is a realistic and safe way to proceed. DiSSCo can easily enable the benefits of applying a Handle-based scheme for Digital Specimens by offering registration and resolution services to clients (data publishers and consumers). Realising this within the DOI ecosystem such that DiSSCo requirements are met (large number of identifiers, tailored metadata scheme, influence to governance) will be developed further in alliance with the DOI Foundation and its RA members. This can be achieved by establishing a new DOI Registration Agency (RA) alongside existing agencies such as

DataCite and Crossref. This RA can be owned, branded and operated by DiSSCo with a scope and mandate different from that of the existing RAs. Working cooperatively as part of the family of DOI Registration Agencies towards establishing a separately owned and operated RA allows maximum flexibility. Full control and accountability can be held by DiSSCo in the medium term with potential to extend internationally to serve the entire global natural science collections community. This will be kept in mind and consultation as development proceeds.

Glossary of terms

Terms and abbreviations used in the present document have the meanings given below.

Multi-Primary Administrator (MPA): An organization, credentialled and authorized to operate and manage (jointly with other MPAs) the Global Handle Registry to allocate and manage derived prefixes from their credential to themselves and to third parties. The MPA and these third parties (such as Registration Agencies) can provide identifier and resolution services (aka local handle services) for handles under the derived prefixes allocated to them.

Registration Agency (RA): An organization authorized by an MPA to provide registration, administration and maintenance services to any legal person/entity wishing to register and maintain PIDs, their references and additional (meta) information.

Handle: A persistent identifier in the Handle System consisting of a Handle prefix and a Handle suffix, separated by a slash (/).

Handle System: A general-purpose global name service run by multiple organisations that allows handles to be resolved and administered securely over the public Internet. The Handle System is a globally distributed implementation of the Identifier and Resolution component of the Digital Object Architecture (DOA).

Global Handle Registry (GHR): A key part of the Handle System that contains records of prefixes allocated to Local Handle Service Providers. A client that queries the GHR will typically learn the network address(es) and certain relevant security information of the Local Handle Services to query for the corresponding Handle/PID record.

Local Handle Service (LHS): A service (organisation and software) for registering, resolving and maintaining PIDs under one or more allocated Handle prefixes.

Local Handle Service Provider (LHSP): An organisation having entered into a Registry Service Agreement with an MPA to provide PID registration and resolution services (local handle services) acts as a Local Handle Service Provider (LHSP).

Persistent Identifier (PID): A persistent identifier is a string (functioning as a symbol/ name) that identifies a digital object. The identifier can be persistently and reliably resolved to digitally actionable meaningful information about the identified digital object.

Funding program

H2020-EU.1.4.1.1. - Developing new world-class research infrastructures

Grant title

DiSSCo Prepare, grant agreement no. 871043

Conflicts of interest

None.

References

- Addink W, Hardisty A (2020) ‘openDS’ – Progress on the New Standard for Digital Specimens. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59338
- Albani Rocchetti G, Armstrong CG, Abeli T, Orsenigo S, Jasper C, Joly S, Bruneau A, Zytaruk M, Vamosi J (2021) Reversing extinction trends: new uses of (old) herbarium specimens to accelerate conservation action on threatened species. New Phytologist https://doi.org/10.1111/nph.17133
- BCoN (2019) Biodiversity Collections Network. Extending U.S. Biodiversity Collections to Promote Research and Education. American Institute of Biological Sciences URL: https://bcon.aibs.org/wp-content/uploads/2019/04/BCoN_March2019_FINAL.pdf
- Casino A, Gödderz K, Raes N, Addink W, Koureas D, Hutson A (2019) DiSSCo Partner Capabilities Survey 2017. Zenodo https://doi.org/10.5281/zenodo.2653707
- Cousijn H, Braukmann R, Fenner M, Ferguson C, van Horik R, Lammey R, Meadows A, Lambert S (2021) Connected Research: The Potential of the PID Graph. Patterns (New York, N.Y.) 2 (1): 100180. https://doi.org/10.1016/j.patter.2020.100180
- Culley T (2013) Why Vouchers Matter in Botanical Research. Applications in Plant Sciences 1 (11). https://doi.org/10.3732/apps.1300076
- Damerow J, Varadharajan C, Boye K, Brodie E, Burrus M, Chadwick KD, Crystal-Ornelas R, Elbashandy H, Alves RE, Ely K, Goldman A, Haberman T, Hendrix V, Kakalia Z, Kemner K, Kersting A, Merino N, O'Brien F, Perzan Z, Robles E, Sorensen P, Stegen J, Walls R, Weisenhorn P, Zavarin M, Agarwal D (2021) Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences. Data Science Journal 20 (1). https://doi.org/10.5334/dsj-2021-011
- Davies N, Deck J, Kansa EC, et al. (2021) Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. GigaScience (accepted, in press).
- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2). https://doi.org/10.3390/publications8020021

- Duckworth WD, Genoways HH, Rose CL (1993) Preserving Natural Science Collections: Chronicle of Our Environmental Heritage. Report of the Conservation and Preservation of Natural Science Collections Project. National Institute for the Conservation of Cultural Property, Washington, DC.. URL: https://digitalcommons.unl.edu/museummammalogy/271/
- European Commission (2013) Directorate-General for Communications Networks, Content and Technology. Digital science in Horizon 2020. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=2124
- European Commission (2018a) Directorate-General for Research and Innovation. Prompting an EOSC in practice. Final report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC). https://doi.org/10.2777/112658
- European Commission (2018b) Directorate-General for Research and Innovation. Turning FAIR into reality. Final report and action plan from the European Commission expert group on FAIR data. https://doi.org/10.2777/1524
- European Commission (2019a) Directorate-General for Research and Innovation. European Open Science Cloud (EOSC) strategic implementation plan. https://doi.org/10.2777/202370
- European Commission (2019b) Reference documents on the EOSC. https://ec.europa.eu/research/openscience/pdf/EOSC_reference_documents_overview.pdf. Accessed on: 2021-2-12.
- European Commission (2019c) EOSC main background documents. https://ec.europa.eu/research/openscience/pdf/EOSC_main_background_documents.pdf. Accessed on: 2021-2-12.
- European Commission (2020) Directorate-General for Research and Innovation. A Persistent Identifier (PID) policy for the European Open Science Cloud. Publications Office of the EU. https://doi.org/10.2777/926037
- European Commission (2021a) Directorate-General for Research and Innovation. PID architecture for the EOSC. Publications Office of the EU. https://doi.org/10.2777/525581
- European Commission (2021b) Directorate-General for Research and Innovation. EOSC interoperability framework. Report from the EOSC Executive Board Working Groups FAIR and Architecture. Publications Office of the EU. https://doi.org/10.2777/620649
- Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, Droege G, Glöckler F, Gödderz K, Groom Q, Hoffmann J, Holleman A, Kempa M, Koivula H, Marhold K, Nicolson N, Smith V, Triebel D (2017) Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database 2017 https://doi.org/10.1093/database/bax003
- Hardisty A (2019) Provisional Data Management Plan for DiSSCo infrastructure. Deliverable D6.6. Zenodo https://doi.org/10.5281/zenodo.3532936
- Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul D, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e54280
- Hui Y (2012) What is a Digital Object? Metaphilosophy 43 (4): 380-395. https://doi.org/10.1111/j.1467-9973.2012.01761.x

- Kahn R, Wilensky R (2006) A framework for distributed digital object services. International Journal on Digital Libraries 6 (2): 115-123. https://doi.org/10.1007/s00799-005-0128-x
- Kallinikos J, Aaltonen A, Marton A (2013) The Ambivalent Ontology of Digital Artifacts. MIS Quarterly 37 (2): 357-370. https://doi.org/10.25300/MISQ/2013/37.2.02
- Khedmatgozar HR, Alipour-Hafezi M (2017) The role of digital identifier systems in the theory of digital objects. International Journal of Information Management 37 (3): 162-165. https://doi.org/10.1016/j.ijinfomgt.2017.01.004
- Lannom L, Koureas D, Hardisty A (2020) FAIR Data and Services in Biodiversity Science and Geoscience. Data Intelligence 2: 122-130. https://doi.org/10.1162/dint_a_00034
- Lehnert K, Klump J, Wyborn L, Ramdeen S (2019) Persistent, Global, Unique: The three key requirements for a trusted identifier system for physical samples. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37334
- Lendemer J, Thiers B, Monfils AK, Zaspel J, Ellwood ER, Bentley A, LeVan K, Bates J, Jennings D, Contreras D, Lagomarsino L, Mabee P, Ford LS, Guralnick R, Gropp RE, Revelez M, Cobb N, Seltmann K, Aime MC (2019) The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. BioScience 70 (1): 23-30. https://doi.org/10.1093/biosci/biz140
- Meadows A, Haak LL, Brown J (2019) Persistent identifiers: the building blocks of the research information infrastructure. Insights 32 (1): 9. https://doi.org/10.1629/uksg.457
- MESRI. (2018) Ministère de l'Enseignment Supérieur de la Recherche et de l'Innovation. Plan National pour la Science Ouverte. France. Accessed 2021-03-02. URL: https://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf
- Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson M (2017) Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use 37 (1): 49-56. https://doi.org/10.3233/isu-170824
- National Academies of Sciences, Engineering, and Medicine (2020) Biological collections: Ensuring critical research and education in the 21st century. The National Academies Press, Washington, DC. USA. https://doi.org/10.17226/25592
- Park S, Zo H, Ciganek A, Lim G (2011) Examining success factors in the adoption of digital object identifier systems. Electronic Commerce Research and Applications 10 (6): 626-636. https://doi.org/10.1016/j.elerap.2011.05.004
- Pearson K, Nelson G, Aronson MJ, Bonnet P, Brenskelle L, Davis C, Denny E, Ellwood E, Herv Goau JMH, Joly A, Lorieul T, Mazer S, Meineke E, Stucky B, Sweeney P, White A, Soltis P (2020) Machine Learning Using Digitized Herbarium Specimens to Advance Phenological Research. BioScience 70 (7): 610-620. https://doi.org/10.1093/biosci/biaa044
- Schindel D, Cook J (2018) The next generation of natural history collections. PLOS Biology 16 (7). https://doi.org/10.1371/journal.pbio.2006125
- Schouppe M, Burgelman JC (2018) Relevance of EOSC and FAIR in the realm of open science and phases of implementing the EOSC. Presented at the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018. URL: http://ceur-ws.org/Vol-2277/paper01.pdf

- SPNHC & ICOM NATHIST (2020) The Role of Natural History Collections in Global Challenges Managing Collections in Crazy Times: Abstracts. VIRTUAL 2020, 8-12 June 2020. URL: https://spnhc.org/wp-content/uploads/2020/06/SPNHC_ICOM-NATHIST-2020-Abstracts.pdf
- Sun S, Lannom L, Boesch B (2003) RFC 3650 Handle System Overview. RFC Editor, USA. https://doi.org/10.17487/RFC3650
- Webster M (Ed.) (2017) The extended specimen: Emerging frontiers in collections-based ornithological research. 1. CRC Press, New York, 252 pp. [ISBN 9781315120454] https://doi.org/10.1201/9781315120454
- Wheeler QD, Knapp S, Stevenson DW, Stevenson J, Blum SD, Boom BM, Borisy GG, Buizer JL, De Carvalho MR, Cibrian A, Donoghue MJ, Doyle V, Gerson EM, Graham CH, Graves P, Graves SJ, Guralnick RP, Hamilton AL, Hanken J, Law W, Lipscomb DL, Lovejoy TE, Miller H, Miller JS, Naeem S, Novacek MJ, Page LM, Platnick NI, Porter-Morgan H, Raven PH, Solis MA, Valdecasas AG, Van Der Leeuw S, Vasco A, Vermeulen N, Vogel J, Walls RL, Wilson EO, Woolley JB (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. Systematics and Biodiversity 10 (1): 1-20. https://doi.org/10.1080/14772000.2012.665095
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

Endnotes

*1 The full discussion can be found here: https://discourse.gbif.org/t/converging-digital-specimens-and-extended-specimens-towards-a-global-specification-for-data-integration/2394.

*2 *toto genere* means "in the whole nature or character."

*3 Collections of Digital Specimens organised in the digital realm

*4 The other key mechanism is a specification for 'open Digital Specimens' (openDS) (Addink and Hardisty 2020).

*5 DOI®, DOI.ORG® and Driven-by-DOI® are trademarks of the International DOI Foundation.

*6 Diverse practices for describing sample-based data in separate parts of the global research community give rise to gaps and challenges when attempting sample-based analyses in multidisciplinary contexts. Cooperative models between stakeholders can thus be fruitful for wider access and reusability (Damerow et al. 2021).

*7 The size of the global community and scale of the undertaking could easily mandate a business model based on a new Multi-Primary Administrator (MPA) approach

(scenarios E1/E2), especially when the scope is extended to cover both the entire World and and digital representations of all kinds of heritage objects. However, such an endeavour, establishing and administering a new top-level prefix/name-segment comes with hurdles and significant long-term obligations. Political, expensive, and difficult to embed within the target community(ies), the MPA approach could only be achieved with full commitment and investment by a global community of actors (i.e., not just DiSSCo) through a lengthy process of consultation and negotiation at international level. With a higher level of maturity and implementation readiness the MPA approach might be a promising option but at present (spring 2021) it remains an idealised desire. This situation can change if the ISO scope of DOI is extended to include other top-level prefixes and rules around obtaining top-level prefixes become more relaxed.

*8 A marketing brochure from the DOI Foundation explains 'driven-by DOI' and gives case examples for Crossref, DataCite and EIDR. Custom characteristics principally include metadata schemas and registration/resolution workflows specific to a sector, but extend also to targeted value added services on top.

*9 See https://www.doi.org/factsheets/DOIHandle.html for more information on how the DOI system uses Handle.net.

*10 See https://arctosdb.org/ and https://handbook.arctosdb.org/how_to/cite-specimens.html.

*11 Although looking at examples where records from other similar CMS have already been registered through Datacite reveals faulty and not very useful metadata, constrained by the DataCite schema.

*12 The GOFA approach has been borrowed from the ARE3NA PID Governance study.

*13 Figures for Crossref revenues obtained from 2018-2019 Annual Report (latest available) published 7 Nov 2019. Figures for DataCite obtained from annual reports 2017 - 2019, which can be found on their governance webpage. Latest annual report for 2019. Figures for EIDR obtained by searching USA Internal Revenue Service (IRS) tax-exempt organization database (TEOS search) using employer identification number (EIN) 27-3656360.

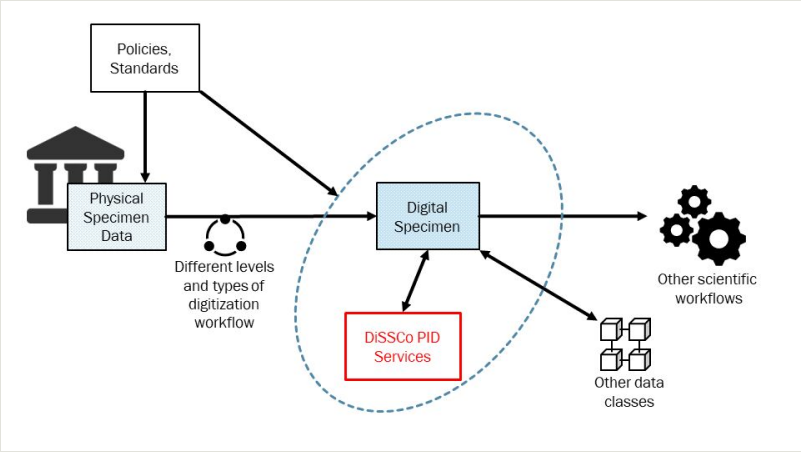

Figure 1.

Digitally transforming collections science with Digital Specimens and persistent identifiers (PID).

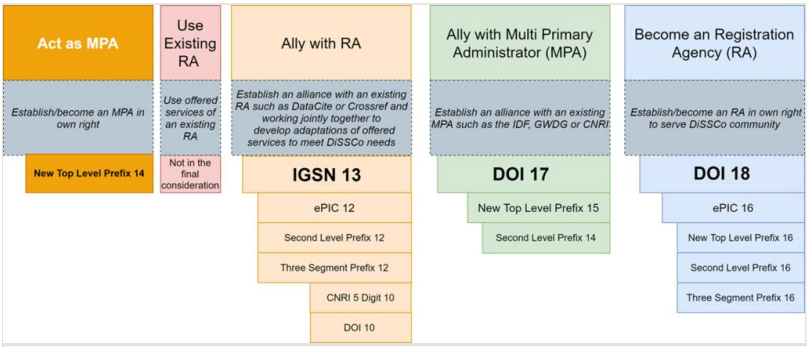

Figure 2.

Scoring of PID scheme optionsKey: MPA = Multi-Primary Administrator, RA = Registration Agency.

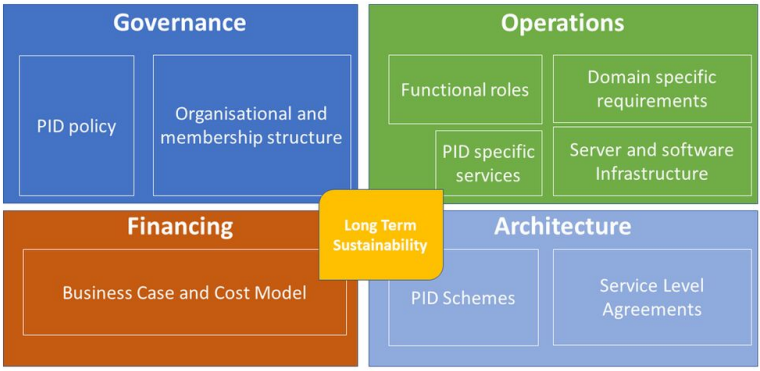

Figure 3.

Governance, operations, financing and architecture (GOFA) together delivering sustainability.

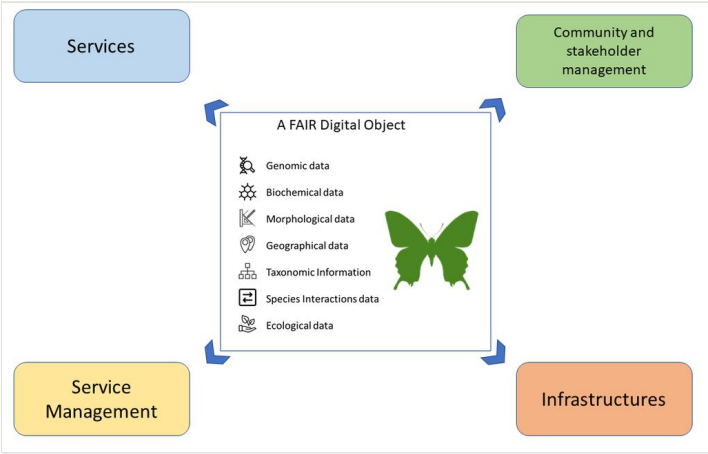

Figure 4.

A PID services model with the essential components in place.

Table 1.

Categories of digital object needing identifiers.

Kind of object	Scenario of use*	PID scheme
Digital Specimen	Internal, external	Topic of the present article
Digital Collection	Internal, external	DOI
Collection Description	Internal, external	DOI
Institution / facility	Internal, external	GRID/ROR#
Loans / visits transactions	Internal only?	t.b.d.†
Annotations / interpretations	Internal, external?	t.b.d.†
Provenance events	Internal, external?	t.b.d.†
Documents	Internal, external	DOI
Persons	Internal, external	ISNI/VIAF/ORCID

* Internal to DiSSCo means PID needs to be resolvable within DiSSCo infrastructure. External to DiSSCo means PID needs to be globally and publicly resolvable.

Sometimes we may need to (internally) reference institutions that do not have a GRID/ROR, e.g., institutions that no longer exist (but their codes are still found in literature and collections), or service providers that are not research organisations.

? Exact scenarios of use need to be studied further to determine whether internal only or both internal and external resolution are necessary.

† The PID type is still to be determined (t.b.d.). Whilst still likely to be selected from one of the Handle System variants, requirements are more 'internal' than 'external' and with lower profile/importance than for Digital Specimens and Digital Collections.

Table 2.

Existing MPAs for DiSSCo alliance.

DOI Foundation (IDF): The IDF is the most well-known MPA organisation, acting to administer Digital Object Identifiers (DOI) or “Dee Oh Eyes” as they are more familiarly known. The IDF is a member organisation governed by a board of directors, with each RA member holding one board seat. The RA members together with the board maintain the focus on scalability, branding, governance and persistence, with the board being responsible for investment decisions. The RAs and the Foundation jointly share responsibility for ensuring that legacy DOIs are maintained in the event of an RA ceasing its membership (which has happened). By definition, the IDF respects and supports the requirements of all its member RAs and claims to be willing and able to implement any changes required by existing RAs and new RAs, as long as such changes do not threaten core principles nor challenge IDF’s ability to guarantee persistence over the long-term.
GWDG: Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen is a service organization working in conjunction with the University of Göttingen and the Max Planck Society as a data and IT service center. It also carries out independent research in the field of computer science and supports scientific research and education. In this latter capacity GWDG provides resilient PID services for its scientific user constituency, and as a credentialled MPA is the leading partner in the European PID Consortium (ePIC). GWDG is presently leading work within the European Open Science Cloud (EOSC) initiative on PID Architectures, services and related requirements.
CNRI: The Corporation for National Research Initiatives Inc., based in Virginia (USA) is a not-for-profit organization formed in 1986 to undertake, foster, and promote research in the public interest, with activities centred around strategic development of network-based information technologies. As the originator of the Digital Object Architecture, the Handle System and other components before turning those over to public ownership via the DONA Foundation, CNRI plays a key role in delivering administrative and technology services in support of the Handle System worldwide. CNRI is both an MPA itself, as well as providing operational services to allow the IDF to perform as an MPA.

Table 3.

Scenarios of PID schemes and potential operational modes for DiSSCo.

Scheme: **DiSSCo modes:**		**DOI (10dot)**	**IGSN**	**ePIC (21dot)**	**Five-digit prefix (CNRI)**	**Two-digit top level prefix**[1]	**Second level prefix**[2]	**Three-segment prefix**	**National-level services**
		A	B	C	D	E	F	G	H
Ally with MPA	1	Possible	Not possible[3]	Possible	Deprecated	Possible[4]	Possible[6]	Not possible	Not possible
Act as MPA	2	Not possible	Not possible	Not possible	Not possible	Possible	Not possible	Not possible	Not possible
Use existing RA	3	Possible[5]	Possible[5]	Possible[5]	Possible[5]	Not possible	Possible[5]	Possible[5]	Not desirable
Ally with RA	4	Possible	Possible	Possible	Possible	Not possible	Possible	Possible	Not desirable
Become an RA	5	Possible	Not possible[7]	Possible	Deprecated	Possible[4]	Possible[6]	Possible	Not desirable

Legend: Ally with MPA – establish an alliance with an existing MPA such as the IDF, GWDG or CNRI.
Act as MPA – establish/become an MPA in community's own right.
Use existing RA – use the offered services of an existing RA without adaptation.
Ally with RA – establish an alliance with an existing RA, working jointly together to adapt offered services to meet DiSSCo needs.
Become an RA – establish an RA in its own right to serve the DiSSCo community.

Notes:
[1] Means a new top-level prefix in addition to those already allocated (10dot, 20dot, 21dot, etc.)
[2] For example, under 20dot, for which CNRI is the MPA.
[3] IGSN uses a 5-digit prefix (10273) for which CNRI is the MPA.
[4] Depends on the existence of an MPA for the prefix. Probably no additional value over the act as MPA option immediately below it.
[5] Restricted metadata capability and generalised services.
[6] Ally with MPA / Become an RA operates in tandem for this PID scheme i.e., both are needed.
[7] DiSSCo assumes an Allocating Agent role rather than a true RA role. IGSN e.V. acts as the RA.

Table 4.

Strength/weakness score of the scenario to support the main DiSSCo requirementsKey: A three-point scale is used 3=strong, 1=weak, 2=in-between, neither strong nor weak.

DiSSCo PID scheme requirement	A1	A3	A4	A5	B3	B4	C1	C3	C4	C5	D3	D4	E1	E2	E5	F1	F3	F4	F5	G3	G4	G5
Scalability	3	1	1	3	1	2	3	1	2	3	1	2	3	3	3	3	1	2	3	1	2	3
Trust	3	1	2	3	1	2	2	2	2	2	1	1	2	2	2	2	1	1	2	2	2	2
Persistence	3	1	1	3	2	2	2	2	2	2	1	1	3	3	3	2	2	2	2	2	2	3
Governance	2	1	1	3	1	2	2	1	2	3	1	1	2	1	3	2	1	2	3	1	2	3
Appropriate identifiers	3	1	2	3	2	2	3	1	2	3	2	2	3	3	3	2	1	2	3	1	1	3
Global adoption	3	3	3	3	3	3	2	2	2	3	2	3	2	2	2	3	2	3	3	3	3	2
Overall	**17**	**8**	**10**	**18**	**10**	**13**	**14**	**9**	**12**	**16**	**8**	**10**	**15**	**14**	**16**	**14**	**8**	**12**	**16**	**10**	**12**	**16**

Legend:
Scalability: Scale for specimens, scale for machines, scale for performance, scale for global use.
Trust: User confidence in the PID scheme, seeing it as appropriate to their needs and trustworthy.
Persistence: Many decades, to more than 100 years.
Governance: An internationally acceptable governance mechanism by stakeholders themselves.
Appropriate identifiers: PIDs appropriate to the digital object type being persistently identified.
Global adoption: Extensible towards a single PID scheme that could be adopted globally.

Table 5.

Strongest scenarios for further evaluation (in ***bold italic***).

Scheme: **DiSSCo modes:**		**DOI (10dot)**	**IGSN**	**ePIC (21dot)**	**Five-digit prefix (CNRI)**	**Two-digit top level prefix**	**Second level prefix**	**Three-segment prefix**	**National-level services**
		A	B	C	D	E	F	G	H
Ally with MPA	1	***Possible (17)***	Not possible	Possible (14)	Deprecated	***Possible (15)***	***Possible (14)***	Not possible	Not possible
Act as MPA	2	Not possible	Not possible	Not possible	Not possible	***Possible (14)***	Not possible	Not possible	Not possible
Use existing RA	3	Possible (8)	Possible (10)	Possible (9)	Possible (8)	Not possible	Possible (8)	Possible (10)	Not desirable
Ally with RA	4	Possible (10)	***Possible (13)***	Possible (12)	Possible (10)	Not possible	Possible (12)	Possible (12)	Not desirable
Become an RA	5	***Possible (18)***	Not possible	***Possible (16)***	Deprecated	Possible (16)	***Possible (16)***	***Possible (16)***	Not desirable

Table 6.

Key performance indicators (KPI) of PID service success.

Desired benefit	KPI definition
Open science: Primary outputs of publicly funded work - the publications and the data associated with and derived from specimens - are publicly findable and accessible in digital format i.e., open. Note, the KPI given here relates to PID services but there can others related to measuring this benefit.	The number of Digital Specimens identified (i.e., having a PID registration) in ratio to the number of specimens digitized, as counted by the Collection Digitization Dashboard. As a percentage.
Reliable referencing: Reliably refer to (i.e., cite) and find the digital equivalent of a specific specimen held in the collection of a specific institution.	Number of citations of Digital Specimens, monthly.
Data accessibility: Reliably access data associated with and/or derived from a specimen.	Rate of increase of PID resolutions, monthly.
Stable, authoritative data delivery: Deliver packages of related scientific information (Digital Specimens) that are reusable and traceable.	Number of DiSSCo participating institutions actively registering PIDs for Digital Specimens (monthly rolling total).
Quality and trust: Strengthened focus on quality and trust in the information handled by the DiSSCo infrastructure.	i) Number of peer-reviewed journal publications referencing Digital Specimens. ii) Number of attributed transactions of work done to improve quality of the scientific information. iii) Number of Digital Specimens where the corresponding physical specimen is not identifiable and traceable back to the institutional collection i.e., for which no digital catalogue record is publicly available.
Added value services: Based on the availability of a growing PID graph of links between multiple Digital Specimens and between Digital Specimens and other data.	i) Number of third-party services available as a result of persistently identifying Digital Specimens (that would not be possible without such identification). ii) Number of published case studies reporting economic, societal and/or environmental impact where such study can be traced back to the availability of services exploiting the PID graph.
Global extension: Serving the needs of the global collections science community. Extension and uptake of the PID scheme outside DiSSCo/Europe.	i) Number of PID registration requests originating outside Europe as a percentage of the total monthly PID registration requests. ii) Number of active non-European PID registrants (active means making PID registration requests in three consecutive months).

Supplementary materials

Suppl. material 1: The type specimen for *Holorchis Castex*; a case example

Authors: Alex Hardisty
Data type: text
Brief description: Supplementary material illustrating how the community ability to effectively link digital representations of voucher specimens with other data types, such as literature, people, genetic sequence information, traits, or even to assert and sustain semantic links between vouchers continues to be seriously hindered by the lack of PIDs and related services.
Download file (22.19 kb)

Suppl. material 2: Estimates of numbers of PIDs needed

Authors: Alex Hardisty
Data type: text
Brief description: This supplementary material provides an estimate of the number of PIDs likely to be needed throughout the DiSSCo lifetime and beyond. Such estimates are necessary for assessing scalability, performance and cost of the alternative PID schemes to be analysed.
Download file (24.42 kb)

Suppl. material 3: Comparison of main roles and responsibilities in the two main categories of administration of a Handle-based PID scheme

Authors: Alex Hardisty
Data type: text
Brief description: Provides a table comparing the main roles and responsibilities in the two main categories of administration of a Handle-based PID scheme.
Download file (22.05 kb)

Suppl. material 4: Dimensions appraisal of the options

Authors: Alex Hardisty
Data type: text
Brief description: This supplementary material details the appraisal for each of the strongest option combinations (Table 8), together with a 'do nothing' option. The appraisal is conducted according to the second step of the two-step approach as explained in the material and methods section of the main article.
Download file (38.97 kb)