Interoperability guideline

Guiding principles for implementing persistent identification and metadata features on research tools to boost interoperability of research data and support sample management workflows

Part of “Enhancing interoperability through the use of PIDs in research platforms“ project
https://eoscfuture-grants.eu/meet-the-grantees/data-and-metadata-interoperability-through-incorporation-pids-research

Vaida Plankytė
Product Design and User Research Lead
Research Space
vaida@researchspace.com

Rory Macneil
Founder and CEO
Research Space
rmacneil@researchspace.com

Xiaoli Chen
FAIR Workflows Project Lead
DataCite
xiaoli.chen@datacite.org
https://orcid.org/0000-0003-0207-2705

31.07.2023

Version 1

The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no 101017536.

Supported by the Research Data Alliance through the RDA Open Calls as part of the EOSC Future project.

Executive summary

In recent years, PIDs have evolved quickly from being an interesting but peripheral resource on the fringes of research data management and scholarly communication to being viewed as a core enabling resource for FAIRification of data. There are an increasing number of PIDs, and existing PIDs are being developed and used more widely. However, PID infrastructure is still immature, incomplete, dispersed, and largely siloed. The critical issue of how PIDs can be incorporated into research tools, and into research workflows, has been largely ignored.

This project was instituted in order to examine in detail how to incorporate PIDs into research tools, using IGSN provided by DataCite, and the RSpace digital research platform, as a case study. The core contributors were DataCite and Research Space, which develops and provides RSpace. Also instrumental in the project work were two research institutions that provided requirements, Rothamsted Research and UiT the Arctic University of Norway, and two tools providers, the Dataverse repository and the new Fieldmark field notebook, with which RSpace has or has a planned integration. They provided valuable input into the design around data and tool interoperability.

Our approach involved extensive interaction with researchers and research administrators from UiT and Rothamsted to understand their workflows and requirements for the use of IGSNs in the context of RSpace. This was an iterative and collaborative process, described in detail in the full report, that made use of user and design research methods such as write-ups, use cases, user stories, diagrams, and feedback to design and implement a solution that accurately reflected and addressed real-life considerations of integrating PIDs within a research workflow.

Throughout the user interaction and subsequent implementation enabling the incorporation of IGSNs into RSpace, the project work was informed by the following design principles:

- Ensure shared understanding of roles and responsibilities that come with the realization of interoperability - particularly regarding metadata creation and management
- Define PID integration goals based on use cases
- Reuse existing metadata frameworks, local and general
- Leverage the open infrastructure to fortify data management workflow
- Work with the disciplinary communities to define best practices

Notwithstanding the very short amount of time available, we were able to implement in RSpace proof-of-concept support for a basic IGSN workflow informed by what we learned from the user research. As part of developing these features, we have integrated with DataCite for handling IGSN registering and publishing actions. This has resulted in a fully working prototype that enables IGSN registration, metadata entry, and publishing all within RSpace Inventory!

RSpace now supports:

- Identifier section present on each sample, subsample, and container
- Register an IGSN for a sample
- Delete a draft IGSN
- Fill in IGSN mandatory metadata fields
- Fill in IGSN recommended fields: subject, description, date, alternate identifier

- Preview landing page
- Publish an IGSN
- Generate public RSpace landing page with metadata
- Re-publish with updated metadata
- Retract a published item
- Publish a retracted item

We have identified the following features as our next steps. We also plan to demonstrate and give our collaborators access to a test server, as this will enhance the quality of the feedback received.

- Add non-IGSN fields to a sample
- Fill in IGSN recommended fields: geolocation, related identifiers
- Make fields mandatory
- Integrate with existing sample template functionality
- Add help text, confirmation dialogs, and documentation

In addition to this work on IGSNs, we plan to take advantage of what we have learned and make RSpace a fully “PID optimized and supporting” resource. This will include:

- Incorporation of support for PIDs for instruments (PIDINST), which we envision will work in a generally similar way to the support developed for IGSNs
- Incorporation of support for the most widely used PIDs in RSpace ELN. Initially, this will include the ability to associate DataCite DOIs, RORs, and RAIDs (when they are available) with records in RSpace ELN.
- Further development of support for relating multiple kinds of PIDs with RSpace resources
- Further support for streamlined use of various kinds of PIDs in workflows involving RSpace and other tools
- Incorporation of support for ePIC handles
- Ongoing, deeper collaboration between Research Space and DataCite, and collaboration between Research Space and other organisations providing handles and PIDs

Table of contents

About the project

The fairly short length of the project required us to narrow the investigation's scope to be feasible. We selected to investigate how RSpace Inventory could integrate with DataCite to support registration, metadata collection, and publishing of the International Generic Sample Numbers (IGSNs) within the context of field sample collection and analysis workflows.

This integration was selected as DataCite had recently implemented support for IGSN ID registration through their systems and DOI API, and RSpace Inventory is a new sample & inventory management tool that is API-based, enabling robust integrations with external tools to be developed rapidly.

What is more, the adoption of IGSNs and persistent identifiers (PIDs) in general has seen heightened interest, thus requiring support in terms of example implementations and guidelines. As a result, the investigation of a sample-focused PID workflow was identified as providing multiple benefits to the research community, as this report contains advice relevant to institutional research data managers, research tool providers, and even development leads.

We are confident that the majority of the advice is more widely applicable to PIDs of any research domain, as many of the implementation and design questions that emerged would remain broadly similar. However, we are not proposing a fully generalizable be-all-end-all guideline, as each PIDs adoption project will have bespoke needs that need to be taken into consideration. We have thus written the guidelines as high-level suggestions that could be adapted to complement such a project.

The contributors

Research Space

Research Space is the developer and provider of RSpace, a digital research platform for Institutional Research Data Management. RSpace features an integrated electronic lab notebook & sample management system that is connected to an ecosystem of tools & services that researchers and data librarians commonly use.

The work was undertaken by Vaida Plankytė, Product Design, and User Research Lead, and supervised by Rory Macneil, CEO. Vaida has a background in Computer Science and works closely with the development team to specify and scope implementation tasks, while Rory has extensive domain knowledge of the emerging conversations and approaches in the Research Data Management and PIDs fields and experience with providing services to institutions with a wide range of structures and needs.

DataCite

DataCite is a global community with a common interest: ensuring that research outputs and resources are openly available and connected so that their reuse can advance knowledge across and between disciplines, now and in the future.

As a community, DataCite makes research more effective by connecting research outputs and resources with metadata—from samples and images to data and preprints. DataCite facilitates the creation and management of persistent identifiers (PIDs), integrates services to improve research workflows, and facilitates the discovery and reuse of research outputs and resources.

This work is supported by Xiaoli Chen, project lead of the Implementing FAIR Workflows project at DataCite. FAIR workflows emphasize the implementation of open scholarly infrastructures, particularly PIDs and associated metadata workflows, in research-supporting tools to support FAIR research practices.

External collaborators

We'd like to thank our various collaborators at FieldMark, Dataverse, Rothamsted Research, and the University of Tromso, the Arctic University of Norway. You kindly provided your time and expertise, which gave us a clear direction for our approach and enhanced the quality of the research tenfold. Thank you!

Understanding PIDs and PID integrations

PIDs and the open scholarly infrastructure

Open scholarly infrastructure encompasses shared resources essential for meeting the needs of authors and readers. Comprising standards, platforms, technologies, policies, and supportive communities, open infrastructure serves as a central facilitator of collaboration. At the core of this infrastructure are persistent identifiers, acting as the "building blocks" that uniquely identify and connect entities within the research ecosystem, including individuals, locations, and objects. The use of persistent identifiers ensures the reliable flow of metadata about organizations, individuals, and objects across systems and platforms, forming the foundation of an open research infrastructure.

In the context of Open Science, "infrastructure" denotes the structures and facilities providing scholarly communication resources and services, including software, that empower the scientific and scholarly community to collect, store, organize, access, share, and evaluate research. Open infrastructure, as exemplified by initiatives like the Principles of Open Scholarly Infrastructure (POSI), encompasses elements of openness, sustainability, community ownership, and interoperability, guiding its development and implementation within the scholarly infrastructure.

For open infrastructure to thrive, organizations endorsing this approach must embrace fundamental principles such as equity, value, trust, interoperability, sustainability, and community governance. Co-creation and active community participation play crucial roles in fostering a sense of "buy-in" and contribute significantly to the success of research infrastructure projects.

Historically, many groups have tended to reinvent tools within their niche instead of building upon existing infrastructures. Nevertheless, some successful counter-examples demonstrate the potential for interconnected infrastructures, like Crossref and Datacite DOIs with ORCID IDs, which enable easy cross-referencing of articles, datasets, and their creators. Recognizing the common need to identify objects in the research enterprise, the development of powerful and often unnoticed infrastructures for identification, storage, metadata, and relationships becomes essential for scholarship. Embracing the benefits of persistent identifiers, such as DOIs, can pave the way for easier creation of discipline-specific services and bridging the gap in describing relationships between various objects and resources.

Open research infrastructures foster collaboration and encourage organizations to approach scholarly research ecosystem improvements from innovative perspectives. By adopting persistent identifiers and integrating them seamlessly into research processes, such as grant applications, manuscript submissions, and repositories, valuable time that was once wasted on administrative tasks can be redirected toward actual research endeavors. Persistent identifiers, as a core element of open research, aim to serve as unique and enduring references for various digital objects, streamlining the scholarly research process.

The fundamentals of PIDs and the core problems PIDs solve

Persistent identifiers (PIDs) play a critical role in scholarly communication by offering unique and enduring references to various entities, such as datasets, papers, and individuals. Unlike uniform resource locators (URLs), PIDs ensure that an entity can always be found, even if it undergoes changes or disappears. This reliability is maintained through machine-readable strings adhering to a defined lexical scheme, exclusively associated with one entity in the world. When an entity is removed, the PID can still direct users to essential information, often presented as a tombstone page, preserving valuable context.

One of the core problems that PIDs solve in scholarly communication is improving interoperability. By creating links between digital entities and enriching them with metadata references to other metadata, PIDs facilitate seamless navigation and integration of diverse research outputs. This interconnectedness enhances the accessibility and discoverability of scholarly information, breaking down barriers between various systems and platforms.

Another crucial aspect is the reusability of data. PIDs enable the association of rich metadata and provenance with digital objects, contributing to the transparency and trustworthiness of research materials. Researchers can easily access contextual information, understand data origins, and assess its reliability, fostering a culture of data sharing and collaboration.

Open PIDs, especially those governed by communities and openly accessible metadata, go a step further in promoting FAIR (Findable, Accessible, Interoperable, and Reusable) data principles. These community-governed PIDs ensure that the attached metadata is openly available under a CC0 license, granting universal access and unrestricted use, and encouraging widespread data sharing and collaboration.

A collaborative effort among key players in the research ecosystem, such as Crossref, DataCite, ORCID, and ROR, provides foundational open infrastructure. They offer various persistent identifiers (Crossref and DataCite DOIs for research outputs, ORCID iDs for individuals) and curate comprehensive, non-proprietary, and accessible metadata. This integrated approach spans borders, disciplines, and time, fostering a connected and inclusive research environment.

PIDs serve as indispensable tools in scholarly communication, overcoming challenges of discoverability, accessibility, and interoperability. They enhance the reusability of data, foster open and collaborative research practices, and contribute to building a robust and FAIR research ecosystem. With the support of community-driven initiatives and open infrastructure providers, PIDs pave the way for a more connected and efficient global research landscape.

Case Study - building interoperability to support IGSN workflows

In the Case Study section, we document the steps taken for the discovery, design, and implementation of a PID integration. Since the goal of the project is to identify considerations when implementing PIDs in a research workflow, the actual *process* of understanding a research workflow and developing PID requirements based on their context of use is also highly relevant to the document.

We hope that an overview of the steps taken in this project will provide other institutions and research tools with a starting template of how to approach this type of work, as the lack of a tried-and-tested list of implementation examples and guidance means that currently, each integration requires extensive research to ensure it supports institutions in their goals.

Additionally, awareness of the points of friction we encountered during our process will hopefully facilitate any further work in the areas of adoption and knowledge building around PID ecosystems.

Process Overview

The process followed is based on user research and user experience design methodology. This enables a user-centric approach that constantly seeks to verify and adapt the design based on feedback from requirement providers and collaborators.

- User research
    - IGSN background research
    - Discussions with IGSN project leads & DataCite
    - Discussions with institutions
    - Detailed write-up of user discussions & research
    - Validation of write-up from users
- Use cases
    - Extended use cases from write-ups
    - Grouping of use cases
    - Validation & adding variations to use cases
    - Identification of responsibilities for RSpace Inventory-type tool
- Implementation
    - Selection of use cases for core IGSN workflow support
    - Prioritising & scoping of use cases for proof-of-concept and minimal working version stages
    - Adapting into tasks for technical and interface development
    - Development
    - User feedback & validation
- Further work
    - Feedback → use cases → development → feedback

In the following sections, we provide detail on the most challenging steps of the project, to enable others to reproduce this approach. We also document the points of friction and our suggestions for minimizing these.

User research

We began by identifying two entities that had expressed interest in adopting IGSNs as part of their research workflow: the Geosciences Department at the Arctic University of Norway (UiT) and the Agricultural Research Institution at Rothamsted Research. Thanks to their engagement, we participated in extensive conversations to gather information on their current use of PIDs, desired outcomes from the adoption of IGSNs, as well as how IGSNs would fit in within their current processes.

Additionally, we complemented this knowledge-gathering process with discussions between RSpace and DataCite, which enhanced the RSpace team's understanding of the core concepts and technical workings of IGSN registration, publishing, and sharing.

We had to ensure we were basing any further work on valid use cases that were provided to us directly by researchers and data librarians, with a valid understanding of the technical underpinnings of IGSNs provided by DataCite. Without holding an accurate understanding of the problem space, we wouldn't be able to design a solution that addressed real-life research pain points.

This is where we encountered our first point of friction: it was challenging to understand and discuss what is possible *technically* and what is the ideal approach *conceptually*, both with institutions and between research tools.

Examples of questions asked (both by institutional and research tool collaborators!) that highlight this issue:

- "Where does the permanent sample page come from, do we as an institution have to provide it?"
- "Does the institution need to become a DataCite service provider?"
- "What is an IGSN, is it a number? Is it a link? Is the link to the institution's page containing the sample metadata, the DataCite Commons page for the item, or is it the doi.org link?

While resources and documentation for understanding IGSNs exist, there is no one place that gathers all essential resources to build up understanding. Some resources are more conceptual and might not answer technical API questions, and others are specific but do not provide an overview of what the moving pieces of integration are.

This issue is exacerbated by the need for parties from different domains to collaborate, which creates a language barrier as not all parties share a background in fieldwork, laboratory work, research data management, PIDs, or software and integration design.

All of these aspects combined caused communication issues, as there was no shared understanding or vocabulary that supported the discovery discussions. We then ran the risk of

building expectations that wouldn't be met if we rapidly progressed to the design or implementation stage.

In an effort to address and minimize this risk, the various conversations were written up and shared with the collaborators in an open document for commenting and consulting with colleagues. Not only is this useful for obvious reasons, but it also enabled us to identify several language and conceptual misunderstandings, as the notes communicated how we understood the requirements from a non-scientist background, which made identifying "wrong" uses of words and simplifications significantly easier once read.

Moving forwards, we ensured that this feedback loop was in place at every further stage of the process and that we were conscious about the language we were using in conversations involving multiple parties. To generate focused feedback, we found that using standard User Research tools helped us refine and present our internal understanding, and facilitated alignment between collaborators.

Use cases

Both institutions described their primary concerns as the ability to record relationships between samples, identify features-of-interest as well as samples, as well as offline data collection.

While their requirements were not identical, due to slightly differing processes and institutional setup, we benefited from being able to verify whether it was possible to generalize the user research into a shared list of requirements that would be more widely applicable to any institution wanting to adopt IGSNs.

While the scope of this project did not allow for a fully interoperable workflow to be implemented across several research tools, we wanted to investigate and include the result of discussions in this report, as they were incredibly beneficial for providing us with the confidence that the resulting working integration would be suitable for interoperating with external tools.

We identified the following tools as the focus of our PIDs interoperability investigation:

- FieldMark: highly customizable offline data collection for fieldwork
- RSpace Inventory: sample management & experimental record linking from RSpace ELN
- DataCite: persistent identifier infrastructure & community-building around PIDs
- Dataverse: open source research data repository software for archiving & sharing

The prototypical metadata flow across these tools is as follows:

- ➔ **FieldMark** (offline sample and metadata collection)
- ➔ **RSpace Inventory** (sample management & experimental record)
    - ◆ **DataCite** API (IGSN registering and publishing)
- ➔ **Dataverse** (archiving & sharing of data)

This selection of tools enables us to model a start-to-finish research workflow, where RSpace Inventory, due to being a tool in the “middle” of the chain, needs to support receiving metadata from a tool and sending that data onwards to the next tool. This leads to an exploration of the requirements for both input and output of interoperable metadata, as well as integration with a PID provider.

As a result, we are able to gather a bigger picture understanding of the points of friction at multiple points of metadata transfer, and for our recommendations to be more applicable to any tool desiring to integrate within a PIDs ecosystem. Note that while our example workflow is focused on fieldwork and IGSNs, we strongly believe the core principles of interoperability to be applicable across domains.

We start by documenting the outcomes of our conversations with FieldMark, DataCite, and Dataverse, which highlight the complexity of a cross-tool integration, and a potential plan for approaching an integration between our tools that would support IGSN interoperability. We then identify RSpace-centric requirements, which form the base for a proof-of-concept implementation discussed in the following chapter.

Interoperability vision

The joint discussion between RSpace, Fieldmark, and DataCite, as well as our conversation with Dataverse, highlighted just how important having a real-time conversation was for the success of the project. We discussed various approaches to interoperating, to answer questions such as:

- Which tools should have the ability to update the DataCite metadata record? What happens if several tools push updates at the same time?
- Does a tool that pushes data onward need to keep track of updates to the metadata made further in the research process?
- How does a researcher know which tool owns the most up-to-date version of the metadata?
- Which tools integrate with DataCite and pull the metadata from their records? Which tools integrate directly with RSpace?
- What is the direction of integration: is it a one-way pushing of metadata toward the next step, or is the metadata ever brought back to a previous tool?
- Should tools push the data forward, or are there cases where the data should be pulled by the tool in the next stage of the process?

The diagram below outlines the workflow that was agreed upon as a viable solution for supporting the metadata flow between our tools: we believe this solution achieves a good compromise, by avoiding the complexity of version management and the need for individual integrations with DataCite. This is achieved by giving RSpace Inventory, i.e. the service provider, sole “ownership” of the IGSN metadata, public landing page, and DataCite actions: registering, publishing, and updating the metadata through the DataCite API.

We also simplify the metadata flow by assuming a linear progression of the metadata through the research process, as FieldMark servers are linked to specific projects and are closed once the project is completed, with the intent of pushing metadata to the next tool after it has been collected.

Once the metadata is in RSpace, any further pushes of metadata, for example, a research data export going into Dataverse, will associate the IGSNs with the export, rather than duplicating the metadata into the export. This avoids the issues linked to duplication and updating of data, as the IGSNs reference the RSpace public landing pages that contain the “source of truth” up-to-date metadata.

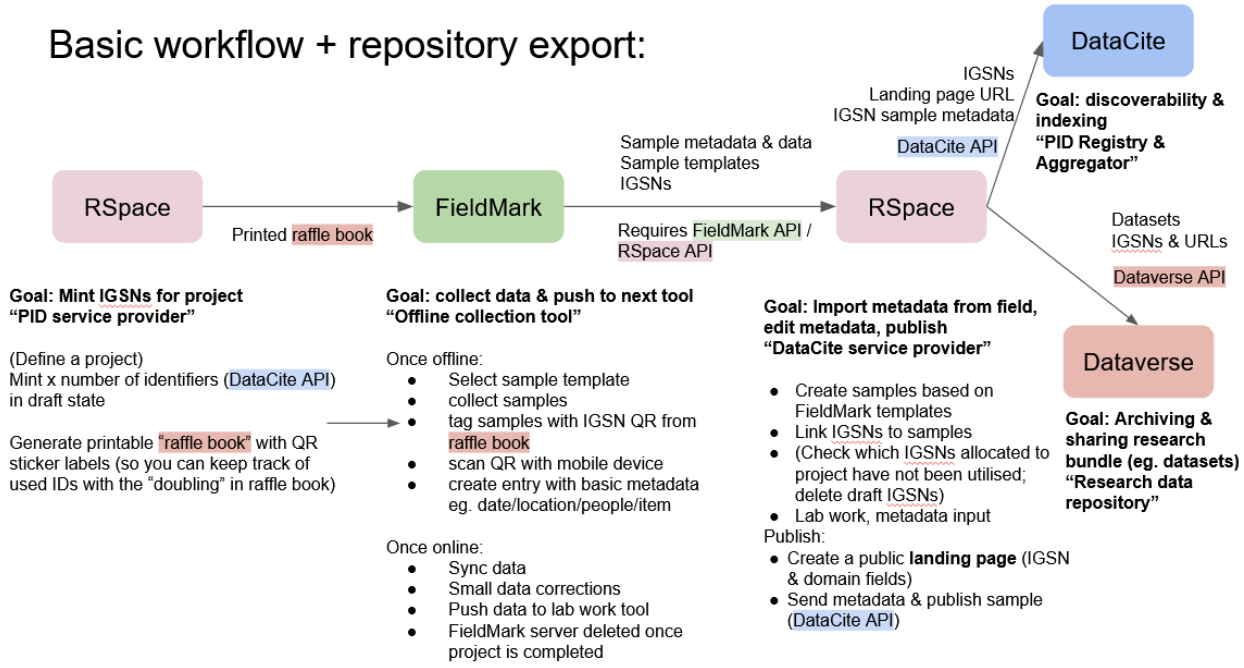


Diagram 1. The basic workflow for using RSpace, Fieldmark, and DataCite, Dataverse integration for sample data management.

We anticipated similar pain points as in the user research stage: it was essential to identify misunderstandings to ensure all parties shared the same understanding of the interoperability plan, especially as the design and development stage would require effort from all parties equally.

We prepared for the discussion by providing an initial diagram to frame our own understanding of what an interoperable workflow utilizing our tools would look like, including several variations that showcased the options we were considering, eg. all tools having access to DataCite for updating metadata. The use of a diagram worked very well in focusing the conversation, especially in providing a visual representation to facilitate thinking.

The discussion identified the need for all tools to adjust their connectivity, that is, enable their APIs to support features specific to IGSN interoperability. For example, FieldMark needed to

support exporting the sample templates and metadata to RSpace, while Dataverse needed to support receiving a list of IGSNs to then display on the archive page. What is more, all tools were needed to ensure that there was no loss of information at a point of transfer, whether this was the metadata content, format, or completeness. This can be exemplified by the requirement for RSpace to support all the metadata field types from FieldMark, to enable full import into RSpace.

As we moved onto the first stage of the integration work, which focused on developing API support for IGSN interoperability features and integration design, we made the decision to focus our project on a specific section of this workflow.

Not only were the API adjustments on our and our collaborators' side going to require time that would delay any integration development beyond the scope of this project, we also identified that we needed to first focus on specifying the RSpace and DataCite-centric part of the workflow, i.e. the "service provider" role, as it is fundamental to enabling interoperability, and would be solely contained within RSpace.

High-level workflow

The following workflow overview highlights the various points of interaction between the selected research tools. It exemplifies a simplified version of the interoperability vision workflow, as it was developed right after our discussions with institutions, which the research tool discussion call expanded on.

We have found that conversations, and conversation write-ups, are better understood by using a combination of simpler and more detailed use cases: the shorter descriptions of a workflow enable easy contextualization of the discussion, while the deeper use cases enable more fine-grained conversations around technical feasibility and concrete user needs.

1. Add sample metadata to **RSpace Inventory** system:
    a. Import from the previous stage, eg. field collection through **FieldMark** or CSV
    b. Manual entry
2. Register an IGSN for a sample (**DataCite** API)
3. Collect measurements, perform experiments, and complete metadata fields
4. Publish:
    a. Create a persistent and public landing page with metadata
    b. Send metadata (IGSN & landing page URL) to **DataCite**
5. **DataCite** handles discoverability through search & filters and displays landing page URLs for more detailed information
6. Push data & metadata to the next stage eg. **Dataverse** archive repository, analysis tool, export to CSV
    a. An experimental record export bundle contains IGSNs and sample metadata within itself
    b. **Dataverse** receives the dataset bundle and IGSNs, dataset page displays landing page links for samples used in the project

Box 1. Sample management workflow using RSpace Inventory, FieldMark, and DataCite

Extended use case list

The multiple-user research write-ups were reviewed to identify the requirements at each step of the research process. By using the *"When _, I want to _ so I can _"* use case format, the exact context and purpose of a requirement are made explicit.

This process generated an extended use cases document, provided in its entirety below. While the exact content of each use case is specific to the workflow, tools, and domain we're investigating, the vastness of the aspects the use cases cover showcases both the sheer complexity and potential for PID integrations.

We hope that further work can continue detailing and expanding on this use cases list, especially in terms of best practices and priorities for usable and reliable metadata input and persistence.

When I'm preparing to put IGSNs in use at my institution…	
I want to…	**So I can**…
Have a clear guide on how to fill in the IGSN metadata in a consistent way to follow my domain's standards	Make my work more discoverable, and there is consistency across domain IGSNs
Be connected to a group of other domain experts to work on discovering and defining best practises and workflows, eg. for IGSN granularity, relationship modelling, metadata for IGSN vs landing page	Ensure the IGSN schema and landing page are maximised in their utility and accuracy in representing samples from my domain, and metadata properties can be suggested to DataCite for inclusion in the schema
Be able to define sample templates in the system	Both to standardise what metadata is collected and recorded, but also to ensure IGSN-specific metadata fields are made available for these samples

When I'm preparing for a field collection trip without online access…	
I want to…	**So I can**…
I want to be able to pre-register a batch of IGSNs in a "draft" state beforehand	Assign unique identifiers when in the field

Print out the QR codes with encoded IGSNs and human-readable IDs in bulk beforehand	Facilitate the tagging process of samples through “raffle books”

When I’m working in the field…	
I want to…	**So I can**…
Record sample metadata offline (eg. using a tool like FAIMS/Fieldmark)	Facilitate field collection and data entry processes
I want to scan IGSN QR labels in the field in order to input basic metadata about the sample	Ensure the item has initial descriptive metadata as soon as it is collected

When I’m back from the field…	
I want to…	**So I can**…
Scan the IGSN QR labels of samples, and automatically create corresponding new sample entries (or select pre-created entries) in the inventory system, with these entries prepopulated with the IGSN and metadata from the offline collection tool	Ensure the sample has a unique ID associated from the start, and data entry workflows are sped up
Bulk-scan items to perform bulk operations/metadata entry	Speed up sample data entry
See the list of pre-registered IGSNs and whether they are in use or associated with a specific field project	See whether any samples were missed in scanning/collecting, or to release the unused IGSNs
Directly register an IGSN when viewing a sample or when selecting a batch of samples	Add a unique identifier to samples outside of a field collection and pre-registration workflow (eg. lab synthesis)

When I’m filling in information about a sample in the inventory system…	
I want to…	**So I can**…

Have support for various metadata field types in the inventory system	Provide a rich set of information on the sample landing page in a readable format Eg. sample images, collection geolocation entries with map display, graph of relationships with other PIDs
Specify IGSN-specific metadata values	Populate the IGSN schema values that are sent to DataCite and appear on the landing page, while respecting the metadata field format restrictions
Specify non-IGSN, domain-specific field values	Share detailed and complete information about the sample on the landing page even if this metadata is not included in the IGSN schema
Add related identifiers and materials to the item Eg. a local ID of the item in institutional database, ORCID ID of contributor, ROR ID for funder, CrossRef for grant, DOI for related journal article, Dataverse DOI for experimental dataset export bundle, IGSNs for parent/child samples	Provide a comprehensive sample metadata entry that uniquely identifies related entities and contextualises the sample, without duplicating information
Indicate relationships between linked items Eg. indicate that a feature-of-interest (collection site) is the origin of a material sample; indicate that a sample originated as a product of two samples in an experiment	Encode a detailed record of the sample's history and enhance the experimental record
Keep the sample unpublished and not indexed publicly even if it has an IGSN assigned	Ensure that it only appears publicly once the metadata entry is fully completed and accurate
Keep some fields private, or provide an anonymised value Eg. Show a general collection location, rather than a specific point that identifiers a collaborating farm	Maintain embargoes and privacy of sensitive information
Access a draft landing page, that I can	Collaborate when preparing for publishing of

share internally within the institution	research materials

When I publish the sample from the inventory system…	
I want to…	**So I can**…
Generate a permanent public landing page that contains both IGSN schema and domain-specific metadata about the sample	Share this sample's information online with anyone, and reference it in my research
Send the IGSN-specific metadata and landing page link to DataCite through API	Ensure DataCite has a record of the sample and its persistent page location
Set the sample as a public findable resource in DataCite	Ensure the sample is available for other researchers to search and find through DataCite Commons or the DataCite API.
I want for the sample metadata to be crawler-friendly Eg. Google Dataset Search? B2FIND?	Ensure the samples are more discoverable across various search systems.

When there are updates to a sample that has already been published…	
I want to…	**So I can**…
Push minor metadata corrections to DataCite and the landing page while retaining the same IGSN Eg. additional measurements	Ensure the public information is as accurate as possible
Take down the landing page and update DataCite to indicate a removal Eg. accidental publishing	Ensure samples that weren't supposed to be published are removed from public access
Create a new version of the sample with a new IGSN, that is linked to the original	Indicate that an item has undergone significant treatment, requiring a

sample Eg. A feature-of-interest has been fertilised, changing its properties and resulting samples	differentiation between the original sample and this new version, so that samples can clearly indicate which version of the sample they have been extracted from
Indicate the status of the sample Eg. sample discarded or destroyed	Clarify to researchers visiting the landing page that the sample does not physically exist anymore, and thus cannot be borrowed

When I want to locate or identify a sample…	
I want to…	**So I can**…
Reference a local ID to find the item if searching internally	Rely on previously used identification systems, rather than just IGSNs
Read the label with a human-readable ID, or scan the QR label into the inventory system	Have piece of mind that the sample can be identified even if the QR code gets damaged
Search DataCite Commons using metadata	Access another institution's sample metadata through their landing pages
Search an institution's landing pages using domain-specific metadata	Perform a detailed search of metadata that might not be part of the IGSN schema, but part of the metadata used by this institution and research field

Core use cases for a prototype

The extended use cases list is complex, covering multiple topics. Based on our experience working on this project so far in terms of ensuring alignment across parties and meeting user needs, and the large possible scope of work as demonstrated by the use cases, we agreed that the only feasible approach was to focus on designing the core workflow first, and expand the functionality based on feedback from collaborators, in an iterative way.

By having a quick feedback loop and building the tool as openly as possible, we would identify design assumptions and adjustments needed early in the process, which is essential when exploring an integration design that hasn't been tried-and-tested and for which templates for design exist.

To select the core use cases, we asked ourselves the question: “What features would be required no matter what approach to integrating research tools and supporting PIDs we select?" We refined the question by focusing on the RSpace-specific part of the workflow, which we had clearer specifications for and was unlikely to change significantly as a result of our conversations with collaborators:

“What are the core set of actions we are *certain* the user would *need* to register, add metadata, and publish an IGSN within RSpace Inventory, using DataCite as a PIDs provider, that would form a *complete* workflow? ”

This narrowed our focus into the following one-line user requirements, alongside the benefits each provides.

I want to…	**Benefits**
Be able to mint an IGSN for a sample	Interoperability, Traceability
Fill in IGSN mandatory metadata values for samples	Completeness, Interoperability
Fill in IGSN recommended metadata values for samples	Completeness, Interoperability
Specify domain-specific field values that are not in the IGSN schema	Completeness
I want to be able to push minor metadata corrections to DataCite and the landing page	Accuracy
I want to be able to take down a landing page and update the DataCite schema to indicate a removal	Accuracy, Privacy
Automatically generate a permanent read-only public landing page for a published sample	Open access
Display the IGSN and domain-specific metadata on the landing page	Open access, Completeness
Automatically send the the IGSN metadata and landing page reference to DataCite	Discoverability, Interoperability

Enable others to search DataCite Commons to discover my samples	Discoverability

These requirements are certainly not a full list that would be sufficient for long-term use and management of PIDs and samples: for example, a major aspect of PIDs workflows is the ability for the institution to manage and standardise metadata input. This would require features such as, for example, sample templates that pre-populate and provide a list of possible values for IGSN fields, as defined by the Community of Practise, the ability to set IGSN fields as mandatory for input, even if DataCite does not require them, and bulk actions such as bulk-registering and bulk-publishing.

This set of use cases is a first step: it covers a self-contained workflow of "register IGSN -> populate metadata -> publish metadata -> maintain metadata", which corresponds to a set of actions performed within RSpace Inventory by a researcher, and interoperates with DataCite through their API.

At this point, we have developed a focused list of requirements that can serve as a foundation for a technical and interface specification. Once this specification is prototyped, the workflow can then be demonstrated to collaborators and institutions for feedback, providing a convincing and thought-provoking proof-of-concept.

Implementation of integration

The development of a proof-of-concept was a core component of the project. By showcasing a visual, working example of how IGSNs can be integrated within an existing research tool, we hope to facilitate conversations around the design of such integrations and enable the discussion to be more concrete and detailed. We are especially looking forward to receiving feedback from the collaborating institutions that provided us with their perspectives, as this will enable us to validate the prototype and enhance the integration design going forward.

We hope that our implementation can be used as a jumping point for new development, where other integration developers start with the proposed core workflow and extend with features that are relevant to their needs.

Finally, we developed the proof-of-concept and wrote detailed specifications to identify more intrinsic issues and unanswered questions that could only be surfaced when needing to make technical and interface design choices regarding each individual element; these are documented at the end of the section.

Features

The features our proof-of-concept currently supports are as follows. As part of developing these features, we have integrated with DataCite for handling IGSN registering and publishing actions. This has resulted in a fully working prototype that enables IGSN registration, metadata entry, and publishing all within RSpace Inventory!

We currently support:

- Identifier section present on each sample, subsample and container
- Register an IGSN for a sample
- Delete a draft IGSN
- Fill in IGSN mandatory metadata fields
- Fill in IGSN recommended fields: subject, description, date, alternate identifier
- Preview landing page
- Publish an IGSN
- Generate public RSpace landing page with metadata
- Re-publish with updated metadata
- Retract a published item
- Publish a retracted item

In terms of future development, we have identified the following features as our next steps. We also plan to demonstrate and give our collaborators access to a test server, as this will enhance the quality of the feedback received.

- Add non-IGSN fields to a sample
- Fill in IGSN recommended fields: geolocation, related identifiers
- Make fields mandatory

- Integrate with existing sample template functionality
- Add help text, confirmation dialogs and documentation

Interface

To showcase the features and design of the integration, we present the interface screens that the user will encounter while going through the workflow. These are screenshots from our development environment, thus this is an accurate representation of the features included in our interactive prototype.

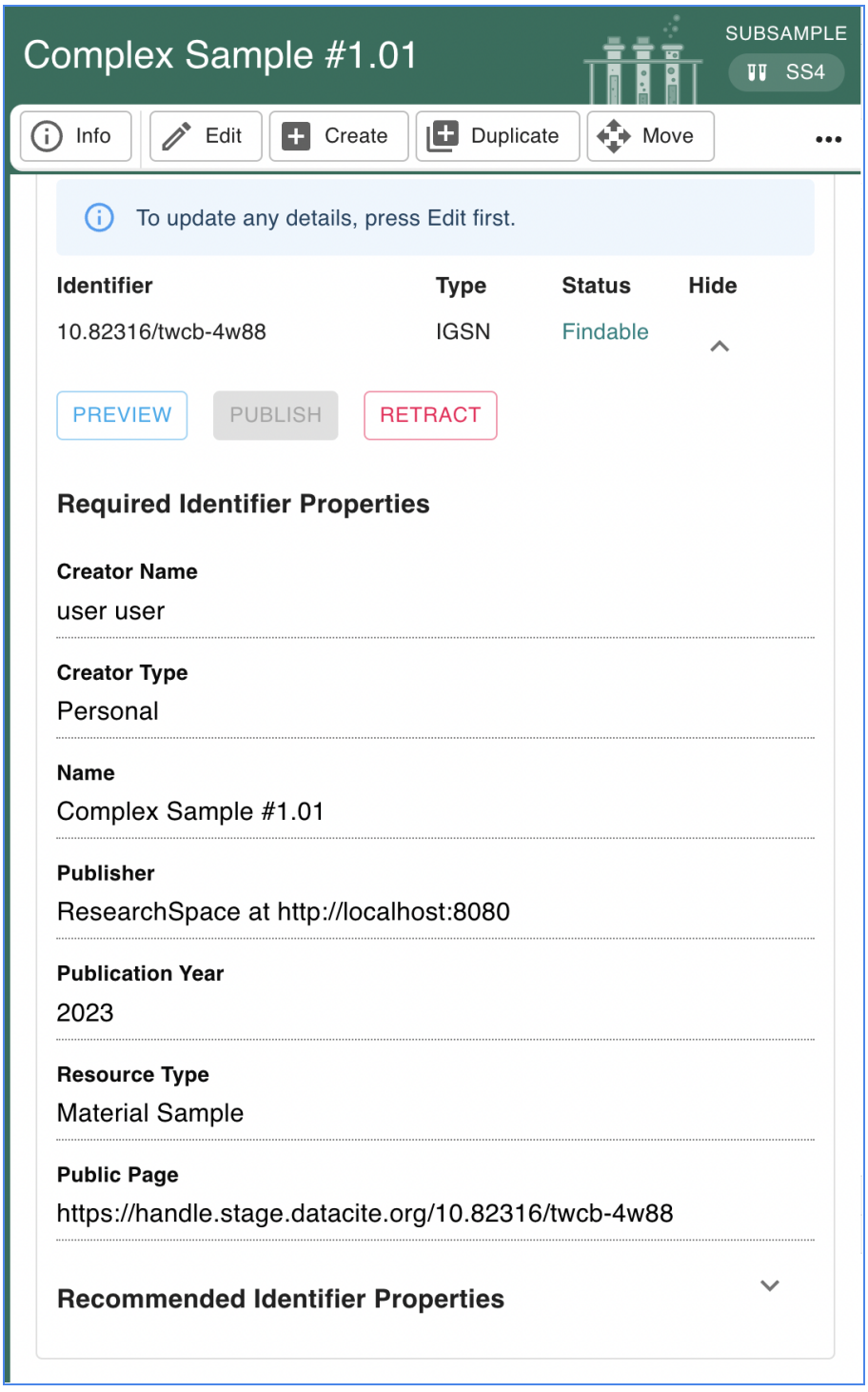


Figure 1. Mandatory metadata entry, IGSN state and actions

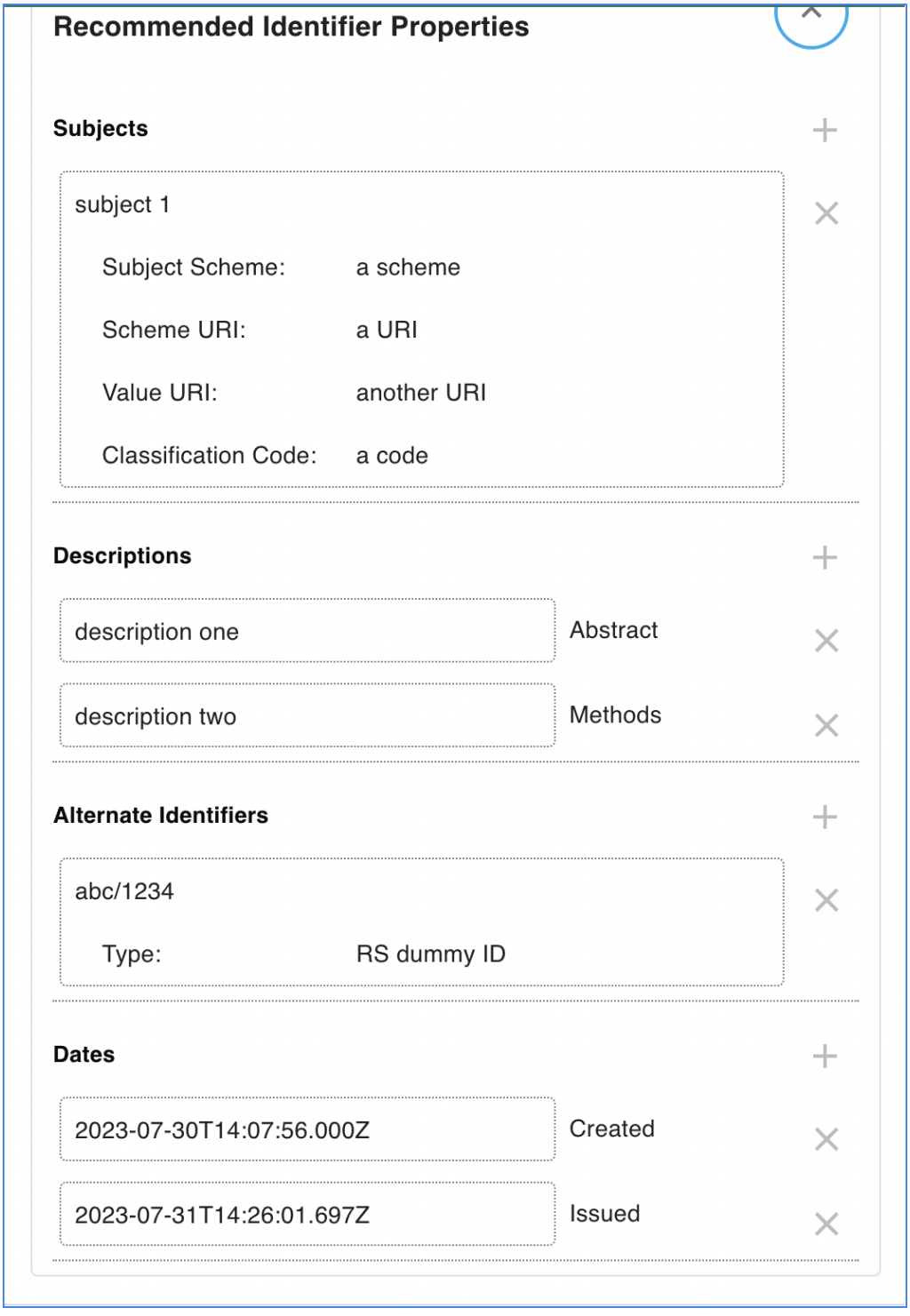


Figure 2. Recommended metadata entry

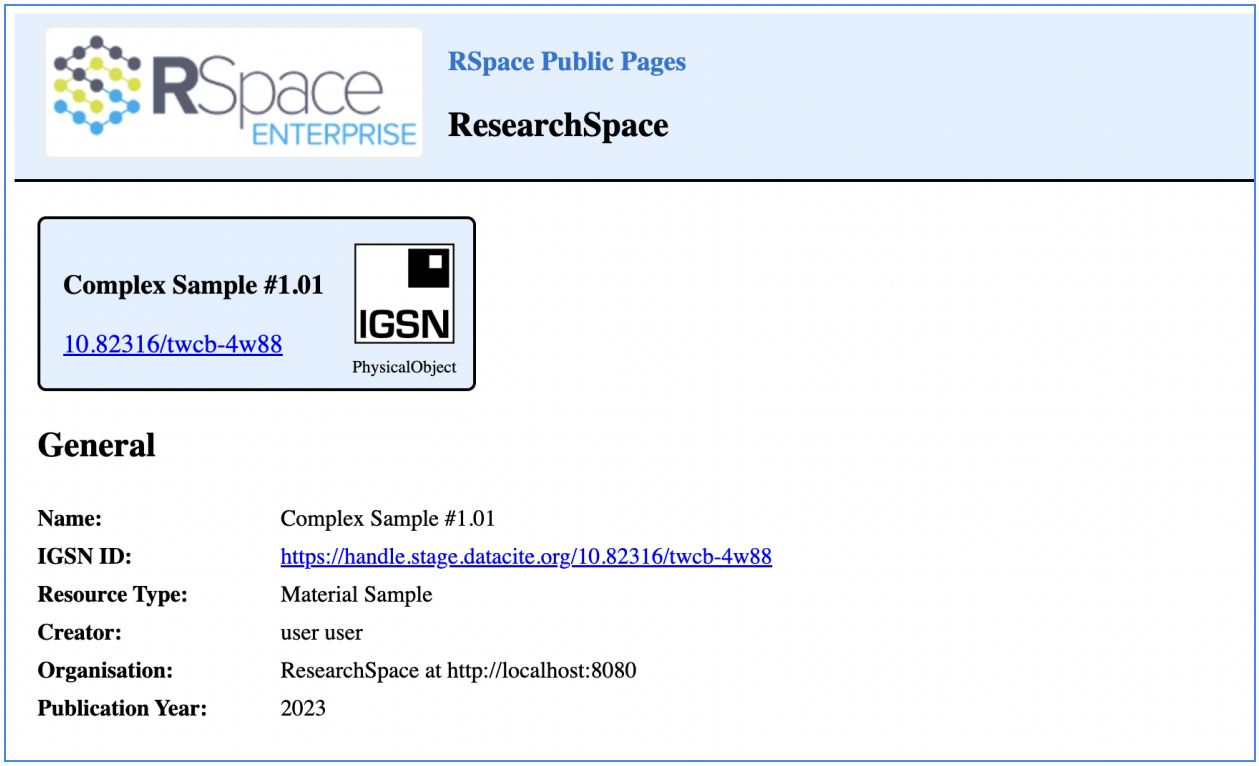
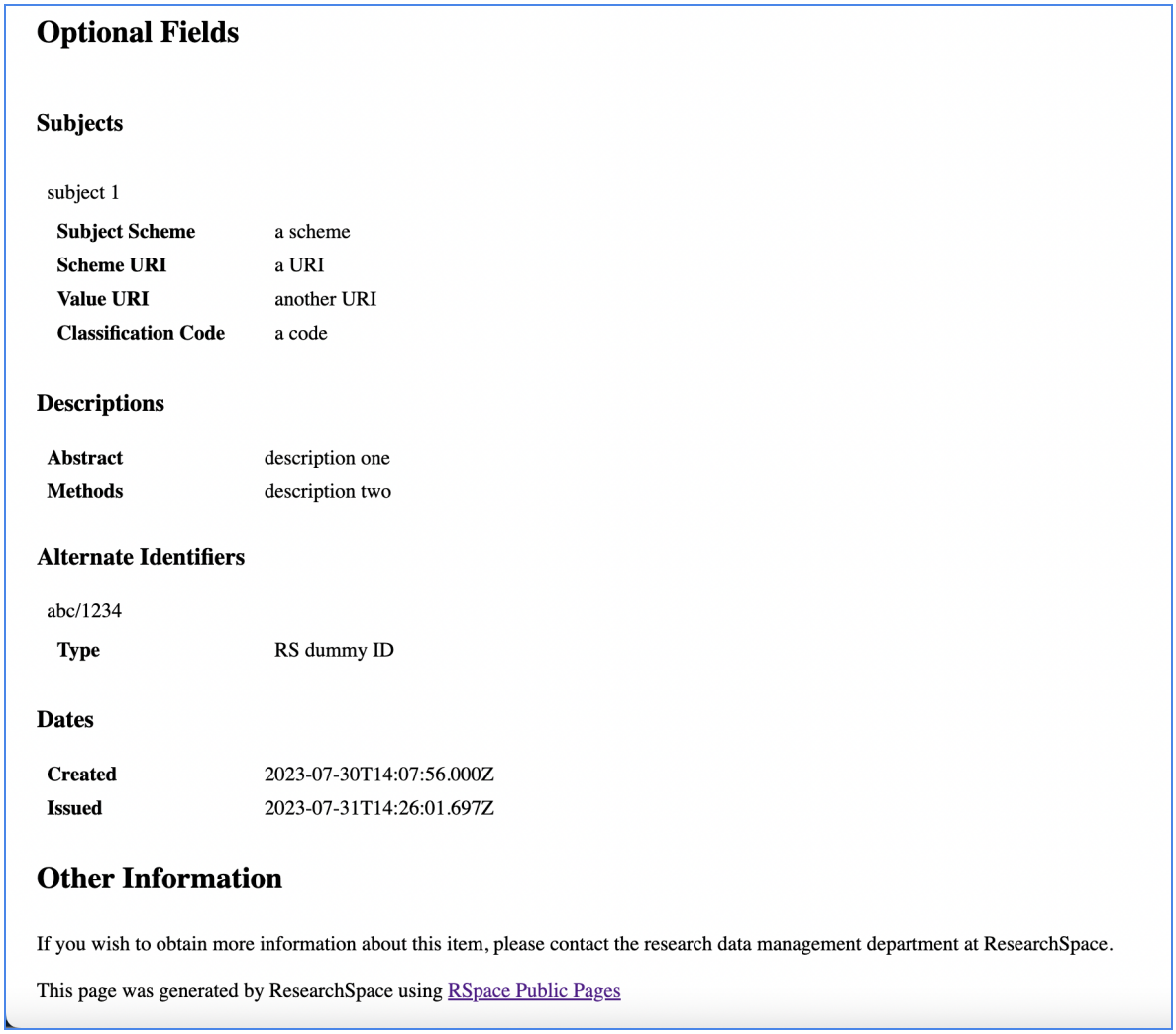


Figure 3. Public landing page 1



Optional Fields

Subjects

subject 1

Subject Scheme	a scheme
Scheme URI	a URI
Value URI	another URI
Classification Code	a code

Descriptions

Abstract	description one
Methods	description two

Alternate Identifiers

abc/1234

Type	RS dummy ID

Dates

Created	2023-07-30T14:07:56.000Z
Issued	2023-07-31T14:26:01.697Z

Other Information

If you wish to obtain more information about this item, please contact the research data management department at ResearchSpace.

This page was generated by ResearchSpace using RSpace Public Pages

Figure 4. Public landing page 2 (optional metadata properties)

Design Principles

We summarize the project outcomes by providing recommendations for any organizations or individuals who are thinking about, currently developing, or want to refresh their PIDs workflow and research tool interoperability.

Ensure shared understanding among all stakeholders. Building interoperability should start with a shared understanding among all parties involved of the roles and responsibilities that come with the realization of interoperability, particularly regarding metadata creation and management. This includes hosting the landing pages of samples/datasets or other types of outputs, as well as curating the metadata associated with the output. Both technical and social integrations are associated with these responsibilities. By integrating with PIDs, the research tool/platform acts as a bridge between local research and research data management workflows and the global open scholarly information community through the interoperable (meta)data model used by various open PID schemes. It is beneficial to all partners—the PID organization, the integrator/tool provider, and the users of research tools—to understand and commit to the open research paradigm, engage in the continued conversation in the community to implement, and improve best practices around research data management.

Define PID integration goals based on use cases. It is important to assess use cases and prioritize integration tasks, as research data management and PIDs workflows can easily generate an overwhelming amount of user requirements for supporting tools. For integrators, it is beneficial to work through the various aspects of the integration and development process in stages, defining the minimum viable product at the start of the process to demonstrate the fundamental mechanism and the core use cases they cater to. This ensures that the product is fit for purpose before moving forward with ease-of-use features that assist the data curators' workflows.

Reuse existing metadata frameworks, local and general. Multiple sets of metadata are usually collected across research tools. For example, domain-specific metadata for sample description and data analysis, archival metadata for object identification and preservation on the institutional level, and metadata elements required for PID registration. The integration of tools should satisfy research data collection, analysis, and management use case requirements, optimize discoverability by adopting existing and community-endorsed metadata standards, and strive for simplified workflows by reusing metadata elements when possible. This requires clarification on who hosts which part of the metadata and how the metadata records can be cross-referenced.

Leverage the open infrastructure to fortify data management workflow. Relatively mature open persistent identifiers exist for researchers, research outputs, research organizations, and more. The metadata generated through the research platform/tool and used for PID registration will be made available beyond the tool itself, for the wider global scholarly community to tap into and build discoverability and boost reusability.

Work with the disciplinary community to define best practices. Continuously engaging with the disciplinary community to learn and take into consideration their evolving research practice, and ensure the interoperability features—either through PIDs integration or direct information exchange with other research tools—remove barriers to an integrated data use and reuse workflow, is key to further adoption of the research tool and the underlying open scholarly infrastructure.

Closing thoughts

This has been an immensely satisfying project to be involved with. The enthusiasm evidenced by DataCite and Research Space enabled a wide-ranging, productive and fruitful exploration and refining of issues, clarification of uncertainties, and in turn progress on development of the concrete support for IGSNs in RSpace described in this report. In addition to the three named contributors, other colleagues from DataCite and Research Space generously contributed their expertise and thoughts to the project work, in a way that materially improved the quality of the outcome.

Neither the report, nor the progress made on incorporating IGSNs into RSpace Inventory, would have been possible without the extended and detailed conversations that took place with representatives from UiT and Rothamsted Research. Their willingness to delve into the nitty gritty details of institutional requirements and researcher workflows was remarkable. Conversations with representatives of Fieldmark and Dataverse were also extremely valuable.

This was a highly ambitious, and quite speculative, project. It was far from certain when the project was proposed that it would result in a set of concrete guidelines for incorporating PIDs into research tools, and yet these guidelines have been developed, as reported above. It seemed even less likely that in the very limited time available it would be possible to gather requirements for, design, and implement support for IGSNs in a complex research tool like RSpace, but again, this was achieved.

We are confident that the fruitful collaboration between DataCite and Research Space that developed during this project will continue and result in opportunities to work together on an ongoing basis and to take part in future projects and in the development of PIDs policy and development. In fact this is already happening, e.g. Xiaoli Chen from DataCite and Rory Macneil from Research Space, as co-chairs of the RDA Working with PIDs in Tools Interest group, have been instrumental in effecting an engagement of that group with the National PIDs Strategies Working Group. The two groups are now exploring ways in which they can work together.

Future work

The project has been instrumental in helping us to better understand the rapidly developing landscape of PIDs, how to make use of current infrastructure, and address limitations and constraints in incorporating PIDs into a research tool in a way that serves the needs of researchers and their institutions.

The difficulties related to user adoption warrant a separate investigation. For example, institutions have to balance standardizing metadata entry to enable an organized, searchable database of work. However, some metadata will not quite exactly fit the provided format, which creates friction for researchers who wish to record their work and may result in inaccuracies.

It is clear that institutions need to continuously collaborate with communities of practice, PID registries, PID providers, data librarians, and researchers to achieve and maintain a balance. A clear process for approaching these conversations and facilitating the research process needs to be defined. Additionally, this process should take place before, in parallel with, and after the design and implementation of an integration, to ensure discovery and alignment between the institution and service provider. Therefore, we recommend further work in this area, as it would provide a “guide for institutions” that is analogous to our “guide for service providers”.

DataCite

DataCite is a community of organizations that share a vision of open and connected research, and relies on the continuous engagement with the members to develop best practices based on open scholarly infrastructure. To help the wider scholarly community benefit from the openly available metadata, DataCite encourages research supporting tools to take advantage of the PID infrastructure to provide interoperability features.

It is evident from the project implementation experience that in order to encourage adoption of PID workflows in research tools, it is key for DataCite to keep investing in its integrator onboarding process by investing in targeted integration support. This could include providing:

- More detailed documentation: The documentation for the DataCite API could be more detailed and include specific examples of how to integrate with the API. This would make it easier for integrators to get started and avoid common errors.
- One-on-one support: DataCite could offer one-on-one support to integrators who are having trouble integrating with the API. This would allow DataCite to troubleshoot specific problems and ensure that integrators are able to successfully integrate with the API.
- Training: DataCite could offer training on how to integrate with the API. This training could be offered online or in person.
- Community support: DataCite could create a community of integrators where they can ask questions and share tips. This would provide integrators with a valuable resource for getting help with integration.

By investing in targeted integration support, DataCite can make it easier for integrators to integrate with the API and get their research tools up and running. This would ultimately lead to more research outputs being registered with DataCite and increased discoverability of research data.

Research Space

In addition to the immediate enhancements to the IGSN functionality described above, we plan to take advantage of what we have learned and make RSpace a fully “PID optimised and supporting” resource.

This will include:

1. Incorporation of support for PIDs for instruments (PIDINST), which we envision will work in a generally similar way to the support developed for IGSNs
2. Incorporation of support for the most widely used PIDs in RSpace ELN. Initially, this will include the ability to associate DataCite DOIs, RORs, and RAIDs (when they are available) with records in RSpace ELN.
3. Further development of support for relating multiple kinds of PIDs with RSpace resources
4. Further support for streamlined use of various kinds of PIDs in workflows involving RSpace and other tools
5. Incorporation of support for ePIC handles
6. Ongoing, deeper collaboration between Research Space and DataCite, and collaboration between Research Space and other organisations providing handles and PIDs