RDA Metadata Standards Directory Working Group: Final Report

Alex Ball Jane Greenberg Keith Jeffery Rebecca Koskela

2016-01-22

Contents

Executive summary

The Metadata Standards Directory Working Group set out to develop a directory that would enable researchers, and those who support them, to discover metadata standards that would be appropriate for documenting their research data, regardless of their academic discipline. It happened that a directory with similar aims had recently been developed independently by the UK Digital Curation Centre (DCC), so the group collaborated with the DCC on developing the directory further to achieve additional goals regarding coverage, ease of maintenance, and sustainability.

The group provided updates and additions to the entries in the DCC directory, and developed a second instance of the directory that could be maintained by the community. Additions and updates to the second instance were and are fed back to the DCC version. The second instance was designed in such a way as to simplify any future development effort, and indeed such development is being taken forward by the Metadata Standards Catalog Working Group.

As well as developing the directory itself, the group also collected use cases that will inform the work of the Metadata Standards Catalog Working Group, the Data in Context Interest Group, and the Metadata Interest Group.

1

Summary of working group activity

1.1 Problem addressed

The value inherent in shared research datasets can only be realized if peer researchers are able to identify, discover, contextualize, interpret and reuse them. They can only do this if the datasets are accompanied by metadata that describes, explains, and associates them with various other entities. When this metadata is missing or deficient, the dataset cannot be used to its full potential, and the scholarly endeavor is poorer as a result.

There are several ways in which the metadata might be problematic. It might not exist at all. It might have to be deduced from unstructured text, a process which is prone to incompleteness and misinterpretation, and requires a level of human attention that is simply not scalable. It might conform to an ad hoc or unsuitable standard, making it inconsistent with the metadata for peer datasets. There may be a proliferation of standards for the given purpose, leading to incompatible silos of data and a dilution of community effort, such that no one standard gets a satisfactory level of support from training materials and tools (Tenopir et al. 2011; Willis, Greenberg, and White 2012).

Such problems can be avoided if researchers endeavor to use existing standards wherever possible, to create local profiles instead of new specifications if the existing standards do not quite meet local needs, and to develop new standards only where there is a definite gap in provision. In order to be able to do this, researchers need access to comprehensive knowledge of the metadata standards that are in use, both within their own field and generally across all fields.

Directories of standards, often broader than metadata standards, have been compiled for a limited number of disciplines. The aim of this Working Group was to set up a directory that is comprehensive across all disciplines, so that it might serve

- researchers not already served by a directory;
- researchers working across disciplines;
- as a backup option for researchers for whom a directory already exists.

1.2 Goals

The Metadata Standards Directory Working Group had three goals:

1. Develop a Metadata Standards Directory that lists standards relevant for research data and is
   - comprehensive, covering all disciplines and most generic applications;
   - easy for anyone to contribute to or update.
2. Define and develop use cases for research metadata, in order to optimize how standards within the directory are arranged and described.
3. Develop a plan for the long-term growth and maintenance of the directory.

1.3 Developing the directory

The original intention was for the working group to develop a new directory from scratch, using previous surveys and subject-specific directories as inspiration. It transpired, however, that in parallel with the establishment of the working group, the UK Digital Curation Centre (DCC) had independently developed its own Disciplinary Metadata Catalogue[1]; this was launched in January 2013. On evaluating the resource and finding that it aligned closely with its own ideals, the working group entered into a collaboration with the DCC to adopt this resource and develop it further (Ball et al. 2014).

In common with the goals of the working group, the DCC directory had comprehensive disciplinary coverage while also catering for general applications, and was being actively maintained and updated. The areas in which the working group hoped to develop the directory were as follows:

1. Deepen the coverage of the directory within the disciplines already covered, extend the coverage to more disciplines, and broaden the utility of the directory globally by including standards and profiles local to areas outside the UK.
2. Find ways in which the directory could be maintained by a wider group of volunteers, and to make the process of contributing entries more transparent.
3. Migrate the service and content to a platform more amenable to future development.

The work of updating the directory was performed by two students, Sean Chen and Cristina Perez, within the School of Information and Library Science at the University of North Carolina, Chapel Hill, under the supervision of the chairs of the working group and with technical assistance from the DCC (Perez 2013). The students conducted a survey in the latter part of 2013 to collect information on disciplinary metadata standards and associated resources. The survey instrument was a web form, a link to which was circulated via the mailing lists of various

[1] http://www.dcc.ac.uk/resources/metadata-standards

stakeholder groups. A total of 41 responses were received, leading to the addition of 14 new standards, 4 new profiles, 13 metadata tools and 19 use cases to the directory, plus updates to a further 18 entries. The survey form[2] remains available for inspection though it is no longer monitored.

The second and third areas were addressed by a reimplementation of the directory using the static site building tool Jekyll[3] and hosted on GitHub Pages. A prototype was developed by Sean Chen and Kate Anne Alderete during the first half of 2014. Further work was conducted to make the directory production-ready in the first quarter of 2015, with editorial work from Adrian Ogletree and a new site theme contributed by Dustin Allen. There are several advantages to this implementation over the original DCC instance:

- New entries and updates can be logged visibly through the GitHub issue and pull request trackers.
- The entries themselves are encoded as plain text YAML[4] files, meaning they are at once easily human readable and machine-interpretable. It also means they can easily be imported into alternative systems.

Further details of this implementation are given under 'The Metadata Standards Directory' below.

1.4 Collecting use cases

The working group developed use cases at its meetings held at Plenary 3 in Dublin and Plenary 4 in Amsterdam.

The use cases at Plenary 3 were focused on how various stakeholders – data custodians, librarians, journal editors, funders – might want to use the directory. The group also developed use cases concerning how the directory might be integrated into various relevant tools (e.g., for data management planning), and whether new services might be possible if the directory could be used as a register of persistent identifiers for metadata standards. Summaries of these use cases are provided under 'Use cases' below.

The use cases at Plenary 4 looked in detail at metadata standards in particular disciplines. The intention was both to understand how researchers might want to search and browse for standards, and to understand how the elements within standards were related to particular applications or tasks. Four use cases were produced by the group, drawn from geospatial engineering, materials science (synchrotron data), humanities and the social sciences, and cultural heritage. These use cases are available from the Metadata Interest Group.

While the use cases did not provide the group with a strong steer about the immediate development of the directory, they helped to shape the thinking of the chairs about how to develop it into a machine-actionable catalog in a future phase.

[2] http://bit.ly/1fToaqd

[3] http://jekyllrb.com/

[4] http://www.yaml.org/

1.5 Achieving sustainability

The DCC directory forms part of the organization's website, which raises some concerns about its sustainability:

- In order to protect the integrity of the main site, administrators must be cautious about granting the requisite rights to edit the directory.
- It is somewhat tricky to export the entries from the system while maintaining referential integrity.
- It would not be possible to redeploy the directory in short order should the DCC site become unavailable.

The development of the GitHub instance of the directory has opened up more possibilities for sustaining the directory into the future:

- Managing the pool of volunteer administrators is considerably easier since granting the requisite rights has no implications beyond the GitHub repository. There are inherent safeguards in the version control system to allow recovery from most problems.
- As mentioned above, exporting the entries from the system is as simple as copying a set of text files.
- Since the site is static, and generated by an open source tool from a set of text files, it would be simple to migrate it rapidly to another host should GitHub Pages cease to be hospitable.

Further details are given under 'Maintenance and retirement plan' below.

2

The Metadata Standards Directory

2.1 Using the directory to look up standards

The Metadata Standards Directory[1] is hosted on GitHub Pages. It supports four types of entries:

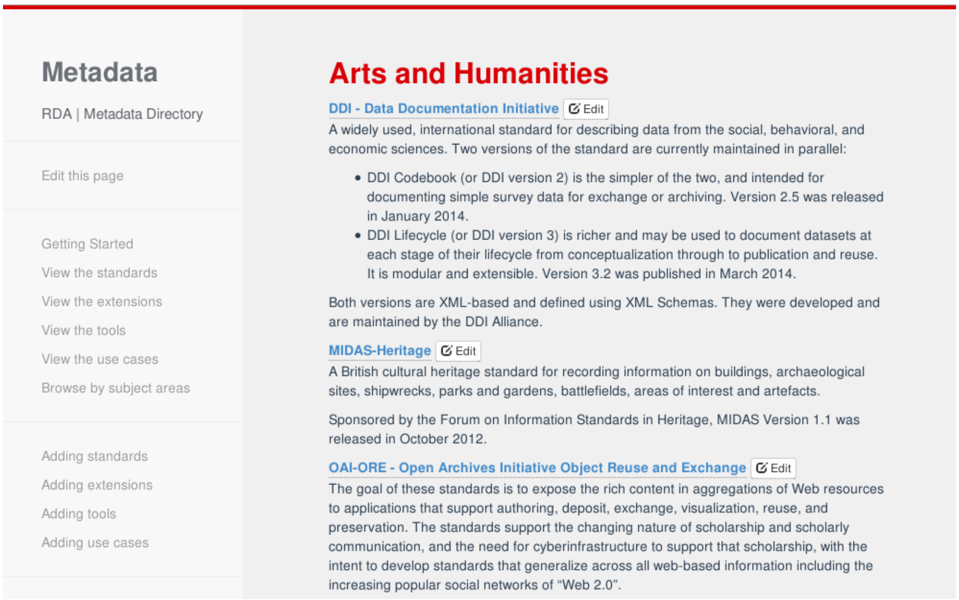
- *standards* are the ‘top level’, independent metadata standards themselves;
- *extensions* are profiles, variations or extensions of the metadata standards;
- *tools* are pieces of software (e.g., libraries, applications, web services) that automate some aspect of using a metadata standard or extension, such as creating an XML serialization from a form interface, or running a validation check;
- *use cases* are examples of organizations and services that use the metadata standards as a regular part of their operation.

The directory provides a separate ‘view’ interface for each type of entry, wherein all the entries of that type are listed. The list is divided into broad subject areas – namely Arts and Humanities, Engineering, Life Sciences, Physical Sciences and Mathematics, Social and Behavioral Sciences, and General Research Data – under which the corresponding entries are listed alphabetically. An entry may be listed under more than one area.

In each case, the entry is represented by a hyperlinked name, an edit button, and a description. In the case of standards, the hyperlink points to a display of the record for the standard; for all other entries, the hyperlink points to the home page for the extension, tool or organization. The edit button is provided so that if users find the description to be erroneous, outdated or incomplete, they can immediately bring up the underlying YAML file for editing. In order to be able to use this facility, users need a (free) GitHub account. Maintainers can make changes directly, while other users need to submit their changes using the pull request mechanism, for which GitHub provides a friendly interface.

Each standard has its own record page. It begins with a recapitulation of the description, followed by a set of summary information:

[1] http://rd-alliance.github.io/metadata-directory/



Arts and Humanities

DDI - Data Documentation Initiative

A widely used, international standard for describing data from the social, behavioral, and economic sciences. Two versions of the standard are currently maintained in parallel:

- DDI Codebook (or DDI version 2) is the simpler of the two, and intended for documenting simple survey data for exchange or archiving. Version 2.5 was released in January 2014.
- DDI Lifecycle (or DDI version 3) is richer and may be used to document datasets at each stage of their lifecycle from conceptualization through to publication and reuse. It is modular and extensible. Version 3.2 was published in March 2014.

Both versions are XML-based and defined using XML Schemas. They were developed and are maintained by the DDI Alliance.

MIDAS-Heritage

A British cultural heritage standard for recording information on buildings, archaeological sites, shipwrecks, parks and gardens, battlefields, areas of interest and artefacts.

Sponsored by the Forum on Information Standards in Heritage, MIDAS Version 1.1 was released in October 2012.

OAI-ORE - Open Archives Initiative Object Reuse and Exchange

The goal of these standards is to expose the rich content in aggregations of Web resources to applications that support authoring, deposit, exchange, visualization, reuse, and preservation. The standards support the changing nature of scholarship and scholarly communication, and the need for cyberinfrastructure to support that scholarship, with the intent to develop standards that generalize across all web-based information including the increasing popular social networks of "Web 2.0".

Figure 2.1: List view of the standards included in the directory.

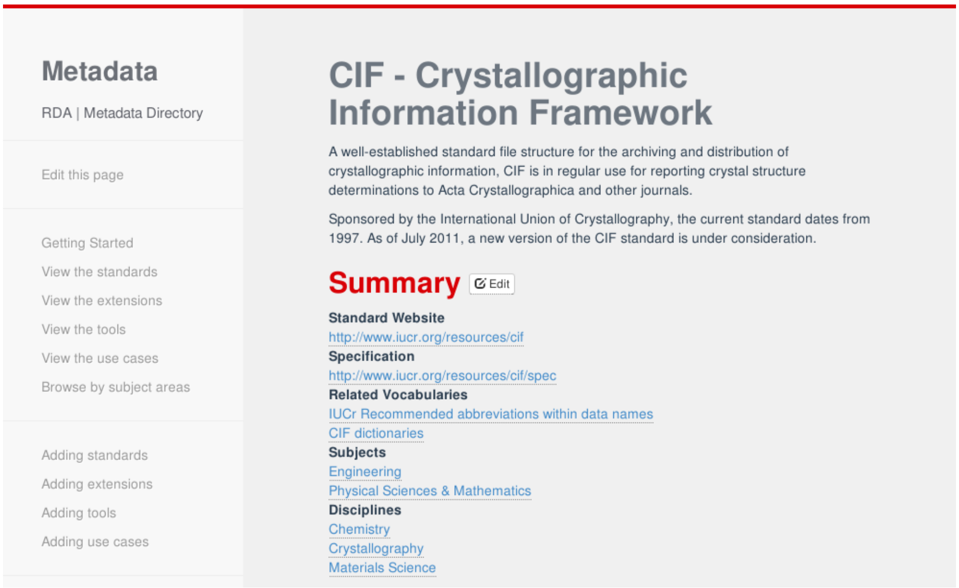
- Links to the full specification for the standard, and to an official website or webpage where more information may be found
- A link to the organization that maintains the standards, if this is distinct from the above
- The current version number and date of last update
- Vocabularies commonly used with the standard
- Mappings to or from other standards
- A contact email address for queries concerning the standard
- Broad subject areas and specific disciplines in which the standard is commonly used

Again, an edit button is provided for quick access to the underlying YAML file.

Lastly, the extensions, tools and use cases related to the standard are listed, again in the form of name, edit button, and description.

In truth, the other entries also have record pages, but since they only provide a name, description and link – and thus do not add value beyond what is provided in the listing – they are not promoted. There is no reason in principle why these other entries may not be expanded and the record pages used, should the user community wish it.

One other browsing pattern is supported: browsing by subject area. An index is provided of the subject areas and disciplines used to classify entries. For each of them, a further index page is provided that lists all the standards, extensions, tools and use cases associated with that subject area or discipline.



CIF - Crystallographic Information Framework

A well-established standard file structure for the archiving and distribution of crystallographic information, CIF is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals.

Sponsored by the International Union of Crystallography, the current standard dates from 1997. As of July 2011, a new version of the CIF standard is under consideration.

Summary

Standard Website
http://www.iucr.org/resources/cif
Specification
http://www.iucr.org/resources/cif/spec
Related Vocabularies
IUCr Recommended abbreviations within data names
CIF dictionaries
Subjects
Engineering
Physical Sciences & Mathematics
Disciplines
Chemistry
Crystallography
Materials Science

Figure 2.2: Beginning of the record page for a metadata standard.

2.2 Developing the source code

The source code for the directory[2] is hosted on GitHub, and versioning is controlled using Git[3].

2.2.1 Branching policy

The source code repository has two branches: *master* and *gh-pages*. The latter is checked periodically by GitHub Pages and is used to regenerate the site. The suggested workflow for developing the directory is as follows (please consult the Git documentation for an explanation of terms):

1. Checkout the *master* branch and pull in any upstream changes.
2. Merge in any changes from the *origin/gh-pages* branch. Note that users who have added or edited files using the links on the directory pages will have made those changes directly to the *gh-pages* branch.
3. Make changes in the *master* branch. It is possible to test the changes by installing a local copy of Jekyll[4] and running it on your local files. Commit and push your changes as and when you are ready.
4. When you are ready to deploy the changes, checkout the *gh-pages* branch, pull in any upstream changes, then merge in the changes from your *master* branch. Finally, push your changes upstream.

[2] http://www.github.com/rd-alliance/metadata-directory
[3] http://git-scm.com/
[4] https://help.github.com/articles/using-jekyll-with-pages/

2.2.2 Arrangement of source code

In the top-level directory, the important files are these:

- README.md contains the high-level documentation for the directory, and is displayed by GitHub when visiting the source code repository in a browser.
- _config.yml contains settings that tell GitHub Pages how to generate the site, and the social media links displayed in the sidebar navigation.
- LICENSE contains the license terms for the code.

Static files related to the site theme are stored in the css, fonts, images, js, and swf directories.

The _layouts directory contains templates for different types of pages, written using the Liquid template language[5]:

- The front page uses the default.html layout.
- The 'Getting started' page uses the about.html layout.
- The layouts for the lists of subjects, standards, extensions, and so on are controlled partly by links.html and partly by the index.md files in the subjects, standards, extensions, tools, and use_cases directories respectively.
- The entry records themselves use the standard.html, extension.html, tool.html, and use_case.html layouts respectively.
- The subject- and discipline-specific lists of entries use the subject.html layout.
- The pages that explain how to add entries use the add.html layout.

The _includes directory contains snippets of content that are injected into the various layouts. The following are of particular interest:

- header.html and footer.html contain the HTML that wraps around the content of all pages. In particular, header.html includes the sidebar navigation and footer.html contains the list of maintainers.
- greeting.html contains the content of the front page.
- standard.yml, extension.yml, tool.yml, and use_case.yml contain, respectively, the YAML templates for the four entry types.

The entries themselves are stored in the subjects, standards, extensions, tools, and use_cases directories. The entries are saved as Markdown files (with extension .md) to ensure they are recognized by Jekyll as pages, but in fact they are empty apart from a block of YAML data. The subject areas and disciplines used to catalog entries are also registered using Markdown files that contain only YAML data; these are kept in the subjects directory.

These last five directories also contain some additional files, namely the aforementioned index.md pages and add.md pages that explain how to add entries.

[5] https://github.com/Shopify/liquid/wiki

3

Use cases

The following are summaries of the use cases discussed at Plenary 3. Note that the majority of them were not within the scope of the Metadata Standards Directory, but suggested how the development of the directory might be taken forward in a future phase.

3.1 Data Providers and Custodians

As a data provider or custodian, I would like to use the Directory…

- …to compile guidelines on the standards that the repositories/archives within my federation should use. I would particularly like to know where mappings exist between standards.
- …to analyze which standards are commonly used by other communities, institutions or archives.
- …to search or browse for metadata standards by what they describe – physical artifacts, video, etc.
- …to compare standards side-by-side, especially to identify commonalities between the standards of different communities.
- …to obtain recommendations of standards to use based on criteria I provide.
- …to search or browse standards by fine-grained tags relating to specialisms within disciplines.
- …to discover the persistent ID for a standard, so I can directly link to it.
- …to discover tools to create metadata based on a standard such as DataCite.

3.2 Librarians

As a librarian, I would like to use the Directory…

- …to obtain assistance in the selection of metadata schemes.

- …to discover ways of adapting a standard to a local need, perhaps by inspecting other profiles of the standard in which I am interested to see how they did it.

3.3 Journal editors and funders

- As a journal editor, I would like to use the Directory to identify what standards exist, and check their maturity and level of support, so I can implement them in my journal's (or publisher's) tools and databases, and include them in the journal's author guidelines.
- As a funder, I would like to use the Directory to find out of which standards we have funded the development, whether they are widely used, whether they have been kept up to date, and whether they be merged into other standards.
- As a journal editor or funder, I would like the Directory to provide me with a clustering of standards that are relevant to my domain.

3.4 Tool developers

As a tool developer, I would like to be able…

- …to submit sample content to the Directory (such as a whole or partial data set) and retrieve a list of metadata standards which could appropriately be used to document that content.
- …to submit a set of field names to the Directory (stripped, perhaps, from an existing metadata record) and retrieve the metadata standard from which they originate.
- …to use tools and services provided by the Directory to map between and align similar ontologies. *(This use case may now be better addressed by the Vocabulary Services Interest Group.)*
- …to use the Directory to convert between schemas automatically.
- …to identify implementors of specific standards using the Directory.
- …to request from the Directory a sample of metadata records adhering to a specific standard.
- …to submit to the Directory the properties (or perhaps certain content) of a data management plan being composed in a tool such as DMPonline or DMPTool, and retrieve a list of appropriate metadata standards to suggest to the user.

3.5 PURLs for standards

As a tool developer, I would like…

- to use the Directory to look up identifiers for metadata standards so I could use them to identify the source and destination of metadata crosswalks. These identifiers should be implemented as persistent URLs such that anyone following the URL would be able to retrieve machine-readable data about the standard.
- to submit a PURL identifier for a metadata standard to the Directory and retrieve the specification for the standard.
- to submit a pair of PURL identifiers for metadata standards to the Directory and retrieve a suggested migration pathway (i.e., a chain of one or more tools or mappings) between them, with some indication of possible loss of information.

4

Adoption

The Metadata Standards Directory is closely tied to the DCC Disciplinary Metadata Catalogue; while the information they contain is not identical, their coverage is the same and they are kept synchronized. The important point is that information that is contributed to the GitHub directory is also pulled into the DCC directory, and in this sense the DCC can be said to have adopted the outputs of the working group.

Similarly, the records in the Metadata Standards Directory have also been imported into the Community Inventory of EarthCube Resources for Geosciences Interoperability[1].

Usage figures are not available for users browsing the GitHub directory, but we can offer some idea of the impact of the information therein by indirect means. The Disciplinary Metadata Catalog is the third most popular part of the DCC website with 27 355 visits in 2015. This accounts for 6% of the traffic to the DCC website and makes it the third most popular section after the events section (14%) and the How-to Guides section (9%).

The DCC directory is recommended as a resource under the heading ‘Identify and use relevant metadata standards’ in the DataONE Best Practices Database. This database attracts 16 500 users and 20 250 sessions per quarter.

The DCC directory has helped some groups in Europe reach decisions to adopt certain metadata standards:

- various groups have adopted DCAT;
- various projects using European Space Agency data have adopted INSPIRE, the European profile of ISO 19115;
- some university biology departments now use Darwin Core;
- some social science departments have moved to using SDMX;
- some universities linked to the UK Data Archive provide support for using DDI.

The GitHub directory has attracted a modest amount of attention from contributors; to date there have been 25 updates or additions suggested through the issue tracker and 21 contributed through the pull request mechanism.

[1] http://earthcube.org/group/cinergi

5

Maintenance and retirement plan

The GitHub directory continues to be maintained by a team of volunteers including Kate Anne Alderete, Alex Ball, Sean Chen, and Adrian Ogletree. Further development of the directory will take place within the Metadata Standards Catalog Working Group, which among other things aims to make the content easier for automated tools to query and act on. Once this has been achieved the resource will be renamed accordingly.

It is possible, even likely, that the Metadata Standards Catalog will not be hosted in the same way as the current Metadata Standards Directory. In which case, the home page of the current directory will be maintained for as long as possible with links directing visitors to use the catalog instead.

The DCC will continue to maintain its Disciplinary Metadata Catalogue in parallel with the GitHub version for the time being. Notifications have been arranged between the DCC and GitHub editors so that each are informed of changes made by the other. It is understood that if and when the forthcoming Metadata Standards Catalog achieves a sufficient level of maturity, the DCC Disciplinary Metadata Catalogue will be retired in favor of it.

References

Ball, Alexander, Sean Chen, Jane Greenberg, Cristina Perez, Keith Jeffery, and Rebecca Koskela. 2014. "Building a Disciplinary Metadata Standards Directory." *International Journal of Digital Curation* 9 (1): 142–151. doi:10.2218/ijdc.v9i1.308[1].

Perez, Cristina I. 2013. "The RDA's Metadata Standards Directory: Information gathering." Unpublished master's paper, University of North Carolina, Chapel Hill.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." *PLoS ONE* 6 (6): e21101. doi:10.1371/journal.pone.0021101[2].

Willis, Craig, Jane Greenberg, and Hollie White. 2012. "Analysis and Synthesis of Metadata Goals for Scientific Data." *Journal of the American Society for Information Science and Technology* 63 (8): 1505–1520. doi:10.1002/asi.22683[3].

[1] http://dx.doi.org/10.2218/ijdc.v9i1.308
[2] http://dx.doi.org/10.1371/journal.pone.0021101
[3] http://dx.doi.org/10.1002/asi.22683