Mainz, 5th April 2024

DSpace-CRIS Praxistreffen Workshop

Slides: https://zenodo.org/doi/10.5281/zenodo.10951235

Participants
- Susanna Mornati   (4Science) susanna.mornati@4science.com
- Andrea Bollini   (4Science) andrea.bollini@4science.com
- Jordan Piščanc   (4Science) jordan.piscanc@4science.com
- Beate Rajski   (TU Hamburg) rajski@tuhh.de
- Philip Münch   (UB Osnabrück) philip.muench@ub.uni-osnabrueck.de
- Sebastian Schick   (UB Rostock) sebastian.schick@uni-rostock.de
- Philipp Rumpf   (Uni Bamberg) philipp.rumpf@uni-bamberg.de
- Daniel Beucke   (SUB Göttingen) beucke@sub.uni-goettingen.de
- Christoph Göpfert   (TU Chemnitz) christoph.goepfert@informatik.tu-chemnitz.de
- Jan Haas   (TU Chemnitz) jan-ingo.haas@informatik.tu-chemnitz.de
- Zaharina Stoynova   (ULB Darmstadt) zaharina.stoynova@tu-darmstadt.de
- Davide Montanaro   (HCU Hamburg) davide.montanaro@hcu-hamburg.de
- Sascha Szott   (HSU Hamburg) <lastname>@hsu-hh.de
- Stephanie Hinrichs   (UB Vechta) stephanie.hinrichs@uni-vechta.de
- Abdulla Schahud   (SuUB Bremen) abdulla.schahud@suub.uni-bremen.de
- Michael Erndt   (Fraunhofer-Forschungsgesellschaft) michael.erndt@irb.fraunhofer.de
- Andreas Sabisch   (FU Berlin) sabisch@ub.fu-berlin.de
- Christina Beckers   (CvO Uni Oldenburg) c.beckers@uol.de

Agenda

1. A quick demo of DSpace-CRIS
2. A look back to the DSpace-CRIS 2023 releases
3. DSpace-CRIS 2024 plan
4. Q&A (from document or Live)

1. A quick demo of DSpace-CRIS

(live)

2. **A look back to the DSpace-CRIS 2023 releases**

2023.01.00 - Released on 9th of June 2023 - Aligned with **DSpace 7.5** tag
Key Enhancements:
- Replaced AddThis as social plugin manager with AddToAny
- Export of the current cris layout configuration as a reusable xls
- Ability to configure the explore sections under hierarchical menu

2023.01.01 - Released on 4th of October 2023
Key Enhancements:
- UI Responsivity improved
- Dynamic edit mode evaluation with custom Filters
- Improved metric donuts in search and categories in statistics
- Datacite schema version update (v. 4.4) switch to the DSpace-CRIS Referer Crosswalk Engine
- Enabled the use of legacy OAI identifiers

2023.02.00 - Released on the 30th November 2023 - Aligned with **DSpace 7.6.1** tag
Key Enhancements:
- Improved ORCID login flow. Now it is possible to complete the login process even when the ORCID account's email is private. (TUHH Contribution)
- Improved ORCID integration with versioned entities. Now only the last versioned item is synchronized with ORCID
- Improved ROR integration. ROR is now exposed to datacite, it is possible to import orgunit from ROR and use the additional information like ROR id and type of orgunit to facilitate the selection of the correct orgunit
- Now the CRIS layout tool support multiple layout for the same entity accordingly to definable custom filters.

2023.02.00 - Key Enhancements
- Improved application performance during SSR. Now requests for pages which contain search results are fetched on CSR only.
- Changed routing behaviour for CRIS item pages. First tab defined by the CRIS layout is no longer used in the item's URL. To maintain backward compatibility redirect to the main page is done automatically.
- Improved managing of metadata security toggle for nested metadata.

2023.02.01 - Released on 27th of December 2023
Key Enhancements
- Added item export format to export items as zip
- Added possibility to restrict export format by groups
- Introduced new metadata to show/hide item's bitstreams using advanced-attachment

2023.02.02 - Released on 9th February 2024
- Created a bootable jar to execute DSpace REST as Spring Boot application (anticipated from DSpace 8)
- Added the internal routing on the counted component and improved its general appearance and behaviour

- In order to have, for ORCID, an external identifier even for orgunits like departments, with this version the orgunit hierarchy is scanned upwards, and the first identifier is used for the disambiguation
- Added support to display an alternative tab/box if the configured one is not accessible by the current user
- Implemented a privacy respectful metrics integration (HSU contribution)
- Improved the post preview on social media when sharing it from the social buttons
- Improved the appearance of subscription emails

2023.02.03 - Released on 29th March 2024
- Major redesign of the virtual metadata and RelatedItemEnhancer behavior to improve performance and allow near-real time update of items due to changes performed on related items
- Added support for pushing the Product and Patent entities to ORCID (donated by floriangantner - University of Bamberg)
- Allowed administrators to always perform PUT/Patches operations over items (in progress or archived)
- Fixed rendering issue with item page with huge/large number of metadata
- Lucky-search configuration now supports also redirect to specific bitstream allowing to preserve URL of individual files moving from other platform to DSpace (i.e. EPrints)
- Fixed issues with type-bind when dc.type is vocabulary controlled and/or provided by a metadata enrichment
- Fixed issue with authority controlled metadata were authority was not saved for repeatable fields
- When linking a page of the repository, the use of a language can now be forced with a URL parameter (e.g lang=it)
- Item thumbnail preview is now disabled in MyDSpace item list
- In items lists, thumbnails are embedded to avoid extra calls and slowdowns
- Added configurable metadata of index-able item bitstreams on Solr
- DOIOrganiser performance improved

DSpace-CRIS version policies

main-cris*: is the new development branch aligned with the development branch of DSpace (8-snapshot until the version 8.0 is out, 9-snapshot later, etc.)*

dspace-cris-2023_02_x*: is the maintenance branch, it will receive at least bug fixes from the corresponding minor release of DSpace (7.6.2, 7.6.3, etc.) plus additional fixes and backward compatible improvement from 4Science and the DSpace-CRIS community*

dspace-cris-2024.01.00*: will be released based on a DSpace 8-snapshot as a preparatory work for the official release. It is NOT recommended to use it in production, it will not receive bug / security fixes*

dspace-cris-2024.02.00*: is expected over the summer, it will be stable and based on the DSpace 8.0 tag*

dspace-cris-2024_02_x*: will be the maintenance branch of the DSpace-CRIS version based on DSpace 8, it will receive at least bug fixes from the corresponding minor release of DSpace(8.1,*

Virtual Metadata in DSpaceCRIS

They have the same purpose than the virtual metadata in the "plain" DSpace: provide the ability to expose information available in a related item (for example the ORCID of the author of the publication) as metadata of the linking publication

Differently than in the plain DSpace they are stored in the database alongside the normal metadata and are expected to be named cris.virtual.<qualifier>

Potentially different implementation can be provided to define the logic used to "enrich" the item, out-of-box a single implementation RelatedItemEnhancer is provided that extract information from the metadata of a linked entity

Plain DSpace uses a realtime read strategy: each time that you request the metadata the virtual metadata are calculated but not stored in the database
- easy to implement
- bad performance when you access the data

DSpace-CRIS uses a WRITE once READ many strategy:
- Data are calculated when the item is created/modified
- Data are stored in alongside normal metadata so that they can be accessed during indexing for search, oai, export, visualization, etc. without any performance penalties
- Both strategies were, until 2023.02.03, subject to data misalignment, if you change the ORCID stored in a person item linked to a publication the virtual metadata is not updated

Two metadata are used for each "virtual"
- cris.virtual.<qualifier> contains the values that you like to see exposed over the linking item
- cris.virtualsource.<qualifier> contains the authorities of the linked items that have contributed to generate the viirtual metadata

The cris.virtualsource.<qualifier> allows us to understand if change made directly to the linking item could impact the virtual metadata, if no there is no need to calculate them again (improvement introduced In 2023.02.02)

Each value of the cris.virtual.<qualifier> is always in pair with a value of cris.virtualsource.<qualifier> , eventually we can have the same authority multiple times in the cris.virtualsource.<qualifier> if we have extracted more information from the linked item

If multiple metadata are used to navigate the relation (for example dc.contributor.author and dc.contributor.editor) we only track and use a related item the first time that we hit it (reduce the computation cost, avoid issues when the same items appear multiple times)

The virtual metadata are generated in a not predictable way, there is no guarantee that the first virtual metadata is related to the authority of the first metadata used to explore the relation

The publication with Authors: Author1, Author2, Author3 can lead to the following virtual :
- cris.virtualsource.<qualifier> = uuid3, uuid3, uuid1, uuid2, uuid2, uuid2
- cris.virtual.<qualifier> = value-extracted-from-author3, value2-extracted-from-author3, value-extracted-from-author1, value-extracted-from-author2, value2-extracted-from-author2, value3-extracted-from-author2

Itemenhancer consumer is responsible to calculate, if necessary, the virtual metadata of an item due to changes performed over the item itself (i.e. edit of the publication metadata). It get the list of related items that have the modified item in a cris.virtualsource.* metadata and track them in a table "queue" named itemupdate_metadata_enhancement.

RelatedItemEnhancerUpdatePoller is executed as a spring periodic task with 5 sec of delay among invocations. It is responsible to process the queue each item one by one to verify that the virtual metadata are still actual and only if changed perform a real item update

4Science planned contributions to the DSpace-CRIS Roadmap 2024

- Improved Audit System: who, what, when with details up to which metadata has been changed from which value to which value
- Support for multilanguage hierarchical metadata: currently metadata within a nested (i.e. Orgunit name of an affiliation, role, etc.) are not shown according to the current UI language
- Mechanism to track check / timestamp outside of metadata: several features need to store a timestamp as a technical metadata on the item to allow incremental processing. This is the case for example for bibliometric update, bibliographic scan, sync procedures. We will introduce a dedicated mechanism outside of metadata to avoid "update" of the item when no strictly needed
- Update the correction and suggestion services: backport to DSpace-CRIS 2023.02.x the result of the Openaire ELD Advanced project merged in DSpace 8 to support multiple providers for corrections and suggestions
- Enhanced Signposting https://signposting.org/ support for DSpace-CRIS
- Administer mode to edit metadata of any DSpace Object (Community, Collection, Bundle, Bitstream, EPerson, Group)
- Expose relation over the SPARQL/RDF endpoint
- Upgrade to the Datacite schema version 4.5
- (Maybe) FAIRiCat https://signposting.org/FAIRiCat/

DSpace-CRIS Roadmap: can you help?

More and more institutions are expressing interest in:
- collecting the contributor Roles (CredIT support)
- Enforce validation and management of hierarchical metadata via an application profile also in the administer mode and over the API (PUT)
- Provide a way for an user to claim its role over a publication, project, etc. (Authorship claim: previously named publication/item claim in DSpace-CRIS 5)
- Collect metrics from the OpenCitation
- Integrate OpenAlex as a live and suggestion provider

4Science add-on modules: https://4science.com/dspace-add-ons/
Community issues and PRs status

- Frontend: https://github.com/4Science/dspace-angular/pulls
- Backend: https://github.com/4Science/DSpace/pulls

From now on PRs to be considered for the maintenance release also require a corresponding PR for the main-cris branch. Only backward-compatible changes are expected with possible exception for critical issues

Meeting Questions and Answers

- *Are there features of DSpace-CRIS 5.x that are not present in version 7.x?*

Discussion:
Authorship claim (previously named Publication / Item Claim), we are looking for contributing institution, we are also open to implement it as an Addon anticipating the upfront investment for the community

Possibility to add objects to existing relations (i.e. add a publication to a project).
 As above
Hierarchical Visualization, it is now provided via a 4Science addon. Once we will recover the development cost it will be released

Network Labs, it is now provided via a 4Science addon. Once we will recover the development cost it will be released

- *does the CKAN integration still work with the latest CRIS version?*

Discussion: this integration was released, it is not updated but institutions can do it themselves, or contribute and we will be happy to take it in charge

- *ROR/departments: can a researcher be related to multiple departments?*

Discussion: yes, but it should be done carefully, for instance specifying the main affiliation

- *What is the current situation regarding the integration of own / institutional S3 storage?*

Discussion: there are institutions using their own storage, it has to be fully compatible with the AWS S3 protocol.

- *how can datasets be managed in DSpace-CRIS*

Discussion: DSpace-CRIS can manage datasets within the CRIS datamodel (as "Products"). The challenge is the upload of large files. DSpace-CRIS offers the bulk upload via excel files, where the location of the bitstreams can be referenced.

Issue: crawlers can create heavy traffic and cause the system to crash. From the server perspective it is not a problem to manage the download, but it has a cost on the infrastructure, so there are measures to prevent it, e.g. download files by chunks.
For video / images, 4Science add-ons can stream the content.

- *challenges of migrating from version 5 to version 7*

Discussion:
Trieste: Users complaining about the changes on the interface, but after some time they appreciated the new system.
Hamburg: it depends on how complicated the system is. A lot of preparation needed for the entities, ORCID integration was a problem at the time.
Bamberg: maybe it was too early because of technical problems.
4Science: version 5 and 7 are two different software platforms with the same name. Architecture has drastically changed, the interface was completely replaced.
Custom data migration procedures can be necessary if data are huge.
Suggestion: perform the migration on a test environment with a devops (reproducible) approach.

- *version 7 or version 8?*

Discussion: 99% of time: version 7 until the end of this year. Version 8 is useful if you really need the new functionalities of version 8. It also depends on the timeline: if you do not need to go live before mid 2025, you can start with version 8.

- *Can COAR Notify help us with the exchange between DSpace and Dataverse? Our use case would be to display the associated research data in our DSpace at publication level.*

Discussion: YES! this is one of the use cases supported out-of-box.
The interesting future of Notify is creating a network of repositories communicating with each other.

- *Does the new virtual metadata code still support transitive relations? (e.g. publication->Author->Institute(OU)->Department(OU), the publication is part of the department bibliography)*

Discussion: Yes, you can configure the virtual metadata on each hop and the queue mechanism ensures that all the data will come properly to the final leaf regardless to the order that updates are performed to the intermediate hop.
You can also bring your own implementation of ItemEnhancer tailored to your specific needs of virtual metadata.

- *Do changes in virtual metadata related fields create a modified push for oai? e.g. OU name change → new modified date for the related publications?*

Discussion: it depends - if you update the item, oai is updated

- *Will it be possible to migrate Audit Log Information from other systems (e.g. EPrints)?*

Discussion: first reaction: quite hard. Eprints has a snapshop of every updated item, represented in xml files, which need to be interpreted and translated into the DSpace format.
DSpace-CRIS in one of the next releases will provide the Audit Log. Migrating from EPrints will provide the latest modification but not the previous ones. The historical xml from EPrints could be stored as a separate file but not integrated in the workflow.

- *Does the deduplication UI now show the actual count of possible duplicates? In 5.10 the count stopped at 100 (might be a Solr restriction?)*

Discussion: in 4Science current Data Quality add-on you have the count and it is paginated. More than two items can be displayed for merge. In-progress items can be deleted and the additional information can be moved to the published item. If the items are both published, the information about the withdrawn item will be preserved and a redirect performed if this item is searched.

- *Is there an easier way to update the layout than the excel file?*

Discussion: the excel file was created for librarians, because it is more friendly than the xml. But a developer could use the REST API to modify the layout. A developer could write a script to turn the excel file into xml and store it (if written by a contributor, it can be added to the main code).
4Science has a plan to implement a User Interface to do that, but it is very expensive and it needs funding. Another interesting development would be to track changes in the database, also looking for funds.

- *Visibility settings are not working*

Discussion: maybe it is not supported by all rendering types, so an issue should be open for investigation.

See you soon at the next DSpace-CRIS Workgroup meeting online on 9th April 2024