

Deliverable D6.4

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	675858	
Deliverable title:	Report on Provenance	
WP No.	6	
Lead Beneficiary:	1: STFC	
WP Title	Report	
Contractual delivery date:	30 July 2018	
Actual delivery date:	30 August 2018	
WP leader:	Tomas Kulhanek	STFC
Contributing partners:	STFC	

Deliverable written by Tomas Kulhanek, Chris Morris

Contents

1	Executive summary.....	3
2	Project objectives.....	3
3	Detailed report on the deliverable.....	4
3.1	Representation of entities	4
3.2	Mechanism to access provenance related information.....	5
3.3	Availability of provenance information	8
3.4	Related work	9
3.4.1	PROV-N grammar and PROV-N editor.....	9
3.4.2	Jupyter notebook.....	10
3.4.3	Secure Export of Settings	11
3.5	CERIF metadata	11
3.6	Domain Specific Metadata	11
3.7	Future Work.....	11

1 Executive summary

As previously reported, each publication in structural biology is derived from a series of processing steps, using data from structural experiments which in turn use samples produced by “wet” laboratory operations. For purposes of data reuse and reproducibility of results, it is desirable to record this whole chain of custody. Current practice falls far short of this. (See project deliverable D7.9.)

The W3C standard PROV-O is designed to support the recording and sharing of such information. We have developed mechanisms for automatically saving this information, storing it, and viewing it. Once the necessary take up is achieved, so that provenance data is recorded, this will allow a user looking at structural results to see the processing steps and, back along the chain of custody, the experimental data.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Provide analysis solutions for the different Structural Biology approaches		
2	Provide automated pipelines to handle multi-technique datasets in an integrative manner	X	
3	Provide integrated data management for single and multi-technique projects, based on existing e-infrastructure		
4	Foster best practices, collaboration and training of end users	X	

3 Detailed report on the deliverable

A publication containing structural biology results is typically based on multiple experiments with data passed through multiple processing steps. For purposes of data reuse and reproducibility of scientific results, it is desirable to record the whole chain of steps, but current practice falls far short of this. West-Life has promoted the use of distributed on-line services for performing and managing the data processing steps, and we have begun to consider how such a virtual environment can be used to record the provenance of the datasets which are eventually made public.

Since we are concerned with the use of on-line services, it is appropriate to consider the available web standards which are being developed in a wider context. We therefore follow the W3C standard PROV-O and the recommendations identified at

<https://www.w3.org/2005/Incubator/prov/wiki/images/0/02/Provenance-XG-Overview.pdf>.

“Broad Recommendations (I): Short Term (1-2 years):

1. There should be a standard way to represent at a minimum three basic provenance entities:
 - a. a handle (URI) to refer to an object (resource)
 - b. a person/entity that the object is attributed to
 - c. a processing step done by a person/entity to an object to create a new object
2. A provenance framework should include a mechanism to access provenance-related information addressed by other standards, such as licensing information of the object, digital signature for the object, digital signature for provenance records
3. A provenance framework should include a standard way for sites to make provenance information about their content available to other parties in a selective manner, and for others to access that provenance information

Broad Recommendations (II): Longer Term (3-5 years)

4. A provenance framework should include a standard way to express the provenance of provenance assertions, as there can be several accounts of provenance and with different granularity and that may possibly conflict
5. A provenance framework should include a representation of provenance that is detailed enough to enable reapplying the process and reproduce it
6. A provenance framework should allow referring to versions of objects as they evolve over time, or to temporal information statements of when the object was created, modified, or accessed. In particular it should provide for a representation of how one version (or parts thereof) was derived from another version (or parts thereof).
7. A provenance framework should include a standard way to represent a procedure which has been enacted
8. A provenance framework should include a way to determine commonality of derivation in two resources"

3.1 Representation of entities

In the context of a West-Life use case and the data life cycle within structural biology, we recommend that a provenance record holds relationships using PROV-O terms and that the W3C PROV-N notation is used. As an example, consider the coordinate file 434159-1.pdb which is the result of docking trials between the proteins gentamicin and the megalin receptor domain CR10. The docking was performed by the Haddock software, using structures obtained from PDB entries 1F8Z, 2LGP, 1AJJ, etc. The work was reported in a published article <https://europepmc.org/articles/PMC3567692> which contains additional information such as the authors of the study and other docking trials. These details should be held in a provenance record associated with the file 434159-1.pdb .

We recommend the following to address the short term recommendation 1 to represent entities:

- a. Dataset file or folder are represented by public/private URI generated by Virtual Folder File picker component with prefix `datafile` and by



- entity** element. Derivation from published structure can be presented by **wasDerivedFrom** element pointing to PDB or UNIPROT entry.
- Person's username or eppn (in case of West-Life SSO) is appropriate for attribution in provenance with prefix **user** pointing to the SSO service endpoint with **prov:type='prov:Person'** for entity definition.
Software or computing workflow is appropriate for attribution in provenance with prefix **tool** pointing to the software or computing workflow entityID or endpoint to the software with **prov:type='prov:SoftwareAgent'** for **entity** definition and referring it in **activity**.
 - Hold set of input parameters for the software tool above as attributes **ex:param1="a"** **ex:param2="b"** of **activity**.

Example provenance record of 434159-1.pdb datafile should look like:

document

```

    prefix datafile <https://portal.west-
life.eu/public_webdav/Yi7z7jBEfnj_KiwXN7qkLHja7mErMRVx_15QXpK+6Ckcm3KbGquw
A+NOD+OgQH7YS1n2ioFms_Ws5ZkLf3O6TaKSjdQuouElCfNoKStH4yQ0ynM7Vn3NSppdrDPygJ
ST/434159-1.pdb>

    prefix dataset <https://portal.west-
life.eu/public_webdav/Yi7z7jBEfnj_KiwXN7qkLHja7mErMRVx_15QXpK+6Ckcm3KbGquw
A+NOD+OgQH7YS1n2ioFms_Ws5ZkLf3O6TaKSjdQuouElCfNoKStH4yQ0ynM7Vn3NSppdrDPygJ
ST/>

    prefix articles <https://europepmc.org/articles/>
    prefix tool <http://www.bonvinlab.org/software/>
    prefix user <https://www.structuralbiology.eu/user>
    prefix pdb <https://www.ebi.ac.uk/pdbe/entry/pdb/>
    entity (datafile:, [prov:label="434159-1.pdb", prov:type="document"])
    entity (dataset:run.cns, [prov:label="haddock
params",prov:type="document"])
    entity (tool:haddock2.2, [prov:label="Haddock 2.2",
prov:type="prov:SoftwareAgent"])
    entity (articles:PMC3567692, [prov:label=" Gentamicin Binds to the
Megalyn Receptor as a Competitive Inhibitor Using the Common Ligand
Binding Motif of Complement Type Repeats",prov:type="document"])
    agent (user:tomas.kulhanek@stfc.ac.uk, [ prov:type="prov:Person" ])
    wasAttributedTo(datafile:, user:tomas.kulhanek@stfc.ac.uk)
    wasDerivedFrom(datafile:, pdb:1F8Z)
    wasDerivedFrom(datafile:, pdb:2LGP)
    wasDerivedFrom(datafile:, pdb:1AJJ)
    activity(wl-tool:haddock2.2, 2018-08-16,2018-08-17,
[ex:param1=dataset:run.cns, ex:param2="b"])
endDocument

```

3.2 Mechanism to access provenance related information

We have implemented the following to address recommendation 2:



- d. Virtual Folder supports the definition of provenance information which links to the entities mentioned in section 3.1. The Provenance metadata is part of the metadata associated with the Virtual Folder file or Virtual Folder directory.

The backend API was enhanced with the optional property `Provenance` and with optional GET query for provenance record only:

```
POST /virtualfolder/api/dataset HTTP/1.1
```

```
Host: 127.0.0.1
```

```
Content-Type: application/json
```

```
Content-Length: length
```

```
{"Id":0,"Owner":"String","Name":"String","Entries":["String"],"Metadata":"String","Provenance":"String"}
```

```
GET /virtualfolder/api/dataset/{id}/provenance HTTP/1.1
```

```
Host: 127.0.0.1
```


```
Content-Type: text/plain
```

```
Content-Length: length
```

```
document
```

```
  //provenance record for dataset with id {id}
```

```
endDocument
```

The frontend UI of the Virtual Folder was enhanced by an additional 'provenance' section in the Metadata tab of the Virtual Folder UI after clicking the small 'badge'  icon next to the file, see Fig. 1. Previous deliverables D6.1 and D5.5 integrated a PDB search and query API into the Virtual Folder file manager, and these remain as other sections in the Metadata tab.

The screenshot displays the West-Life File Manager interface. At the top, there is a navigation bar with links: Virtual Folder, Home, Services, Support, News, About, Cloud, Developers, Contact, and Introduction. Below this, the 'File Manager' window is open, showing a 'File List' tab with a directory tree. The 'Metadata' tab is selected, showing details for the file '434159-1.pdb'. The metadata section includes fields for name, extension, path, size, date, local url, and public url. Below this, there is a 'Free text metadata' section and a 'Related PDB, Uniprot, ... items to add' section. A 'Generate' button is present, and a dropdown menu is open, showing 'Ligand (2)' and 'PDB EMDB id (31)'.

Figure 1. Metadata tab available for each file or directory contains generic section, entries section and provenance section which is empty by default.

If provenance section is empty (no record is present for a particular file or directory) the following actions are allowed:

- e. Generate a record in PROV-N notation for the selected file in Virtual Folder, and continue to edit it within the UI
- f. Save the record within the metadata of the Virtual Folder

After the “Generate” button is pressed a generated provenance document can be viewed and edited as seen in Fig.2.

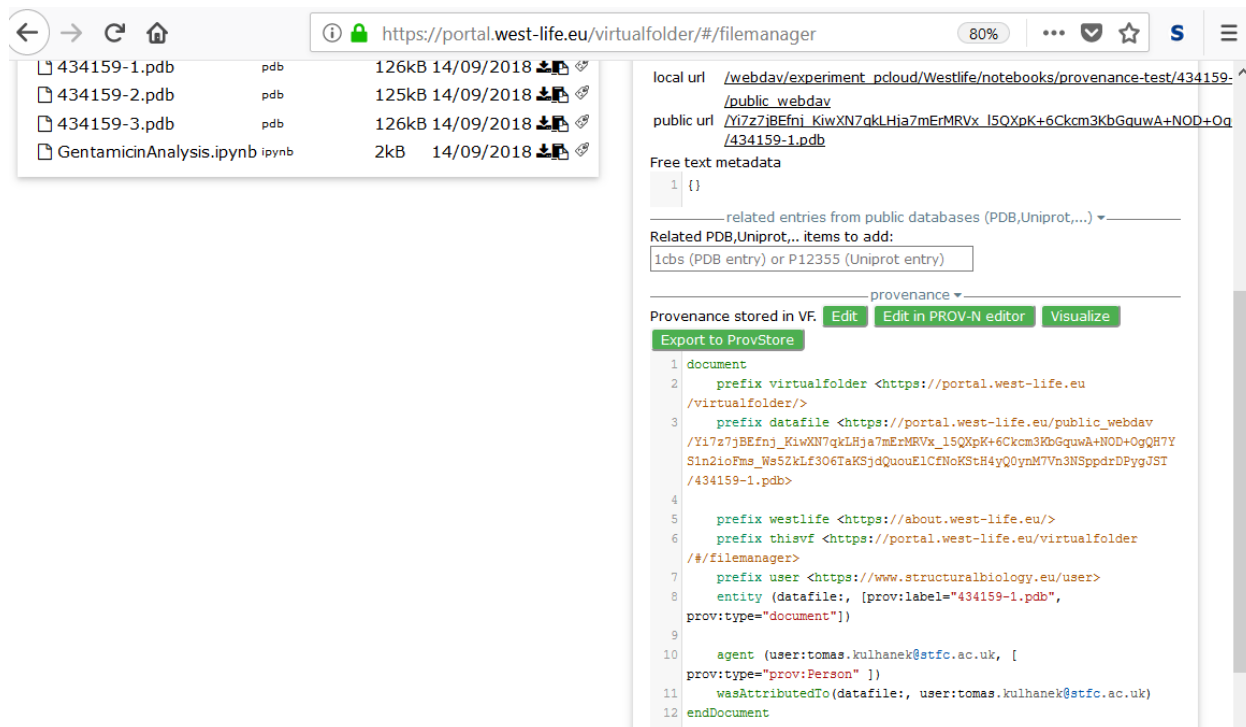


Figure 2 Provenance section of Metadata tab after “Generate” button was pressed. The generated provenance record is in W3C PROV-N standard notation.

Note that the auto-generated provenance record is basic, and misses the connection to 1) software, 2) software parameters, 3) original article (compare with the provenance record shown in section 3.1). Therefore a button “Edit” allows the user to update the record directly in the provenance section, with syntax highlighting. However, proper syntax checking is only available when pressing “Edit in PROV-N editor” button, which opens a popup window with a more complex editor, see Fig. 5. This information is to be added manually, however, should be considered for autocompletion based on usual context for structural biology, see section “Future work”.

3.3 Availability of provenance information

We have implemented the following to address recommendation 3. The Virtual Folder allows researchers to:

- g. Export the provenance document into King’s College London Provenance Store <https://openprovenance.org/>. This is available via the ‘Export to ProvStore’ button, see Fig 2..
- h. Provide HTTP header (PROV-AQ) for resources which have associated provenance record. When a provenance exists for a specific dataset or document, the appropriate record is made within Virtual Folder Web Server Apache. When such resource is checked using any HTTP request next time a Link header is appended.

- i. Provide Web widget to detect and show provenance if link is presented in HTTP header. See Fig.3

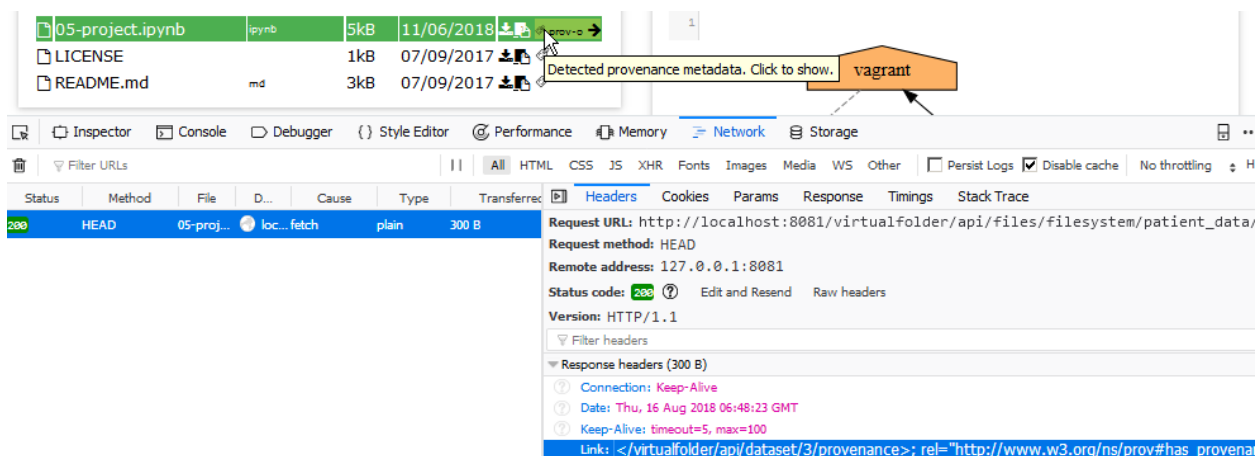


Figure 3. When HTTP Header Link is presented in the response header, the existing provenance link is presented as modified icon with prov-o label and direct link to provenance data

The widget is part of the Virtual Folder File Manager UI. There is also an independent widget which accepts the URL of a provenance document as a URL hash parameter e.g.:
http://localhost:8081/virtualfolder/prov-n-widget/#url=http://localhost:8081/virtualfolder/api/dataset/2/provenance as seen in Fig.4.

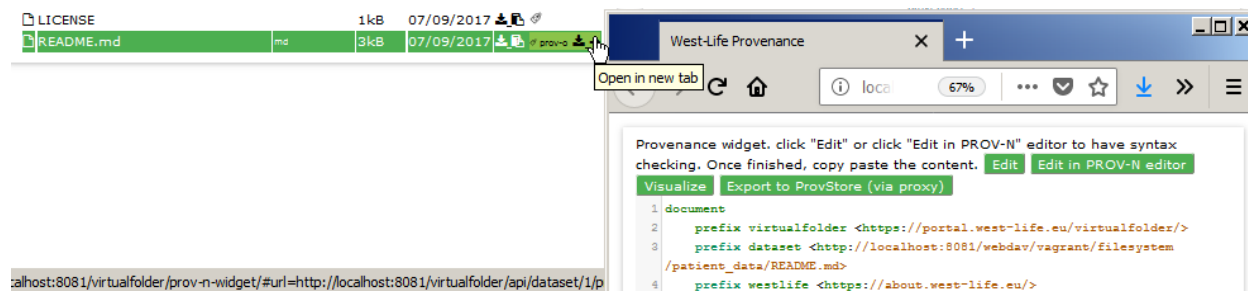


Figure 4. Provenance widget, opens provenance link in separate tab with editor and visualization

User's guide <https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/users-guide/metadata>
and Developer's guide <https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/developers-guide/metadata-and-api/dataset-metadata-and-api>

3.4 Related work

3.4.1 PROV-N grammar and PROV-N editor

During development, it was recognized that there exists a PROV-N editor with some syntax highlighting and snippets at <https://provenance.ecs.soton.ac.uk/tools/editor/>. However, no instant syntax checking per formally defined grammar is done by the mentioned web based editor, thus, a new formal grammar of PROV-N standard for ANTLR v 4 was defined. Integration with browser-based editor ACE was developed as a new web widget PROV-N-EDITOR <https://github.com/h2020-westlife-eu/prov-n-editor> available online at <https://h2020-westlife-eu.github.io/prov-n-editor/>. PROV-N editor is integrated into Virtual Folder UI using pop-up window and cross-document messaging API. See Fig.5.

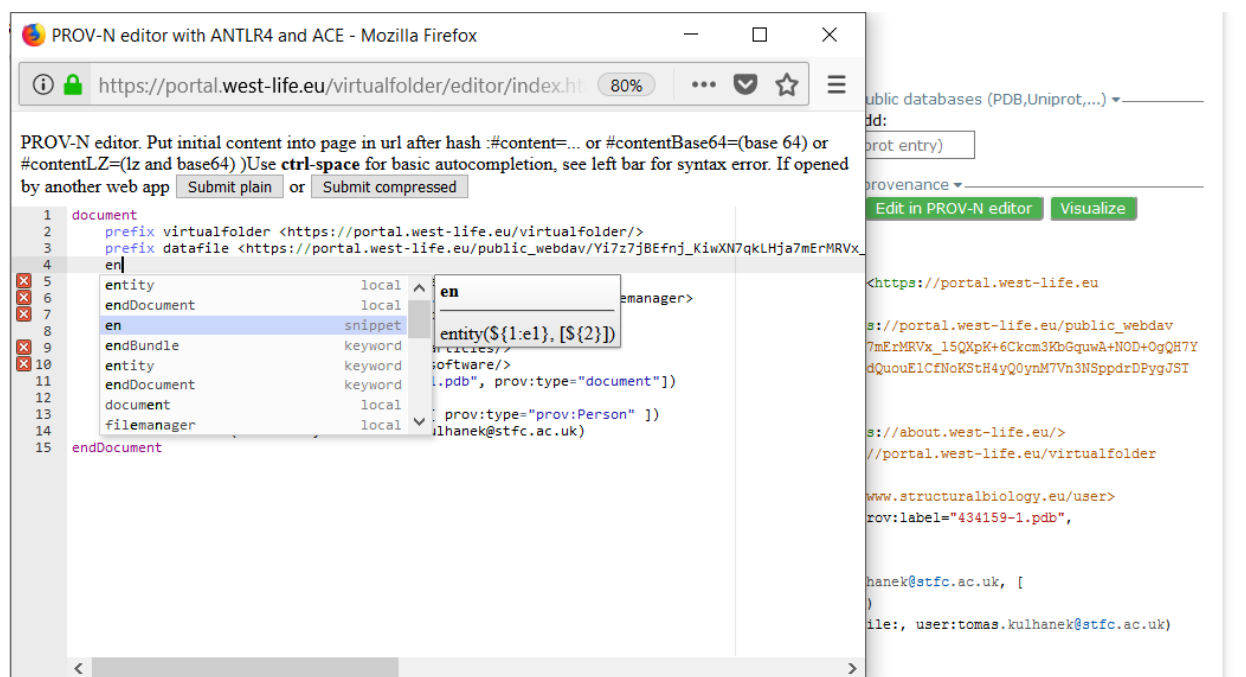


Figure 5 pop-up of PROV-N editor with syntax highlighting, syntax checking, snippet and syntax autocompletion.

The ANTLR v4 grammar was committed into <https://github.com/antlr/grammars-v4/tree/master/prov-n> and authors of ProvStore were notified about the grammar and the new PROV-N editor.

3.4.2 Jupyter notebook

Jupyter notebook is a [web-based interactive](#) computational environment for creating interactive documents in various languages. We prepared optional installation of Jupyter notebook and Python 3 and R, with libraries for data science, structural biology and provenance standard PROV-O available for local deployment of Virtual Folder via CernVM-FS. User's guide <https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/users-guide/related-applications/jupyter-notebook>, Developer's guide <https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/developers-guide/related-application-and-services>

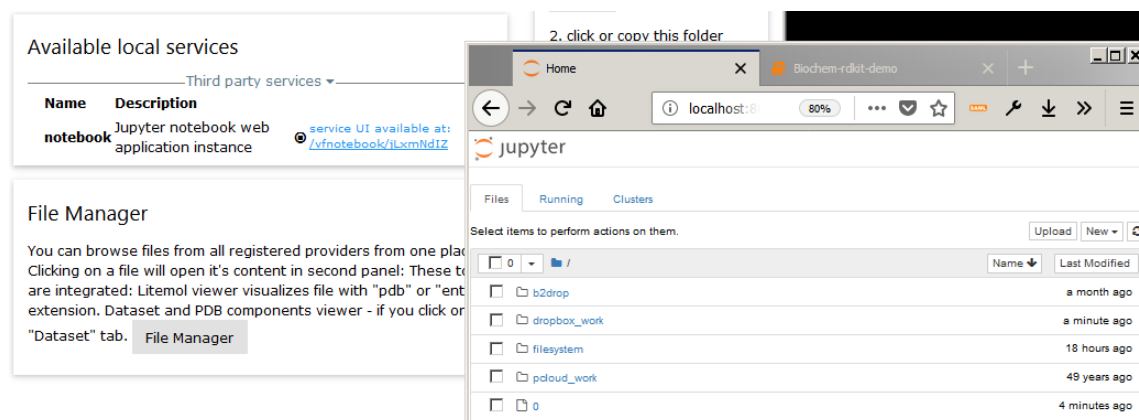


Figure 6 Instantiating Jupyter notebook and open window in the location provided by Virtual Folder UI. Storage locations configured with

Virtual Folder are visible and available for Jupyter app as directories.

3.4.3 Secure Export of Settings

In order to connect a local deployment of the Virtual Folder with a public portal instance, a new feature to export/import settings was implemented using one-time generated asymmetric 2048 bytes long RSA key and random symmetric AES key. It allows to copy Virtual Folder settings from one instance (public portal) to another instance (private deployment).

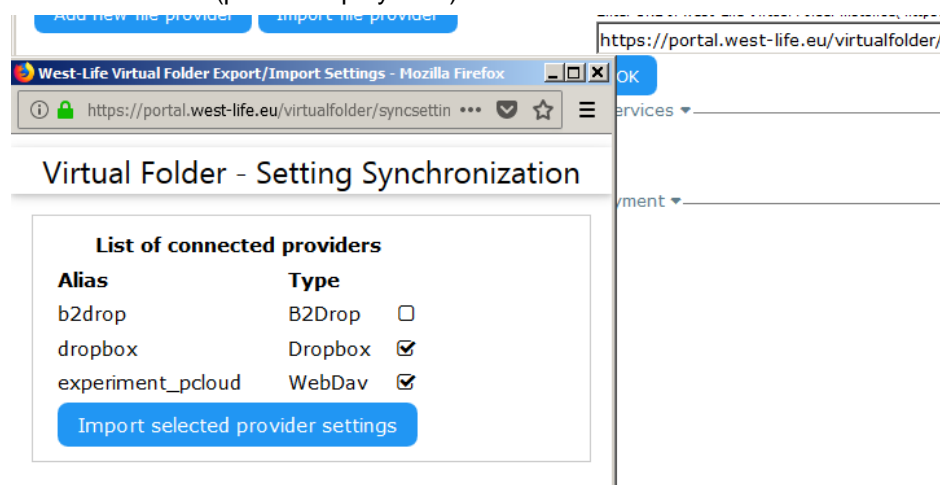


Figure 7. Virtual Folder – Importing setting from another Virtual Folder instance.

There is a user guide at <https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/users-guide/settings/import-settings-from-another-virtual-folder> Developer's guide at <https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/developers-guide/import-export-settings-api>

3.5 CERIF metadata

The CERIF standard is an XML based vocabulary to exchange information about Persons (researchers/authors), Organisational Units (institutions, institutes, etc...), Projects, Publications, Datasets and Services.

A dataset metadata can be exported into CERIF format using the dataset metadata URI and appending a '/cerif' suffix. This request returns metadata in CERIF format, which can be consumed by any other service. Datastores implementing CERIF are dspace, and dataverse. In order to consume and harvest CERIF metadata, we recommend that developers select existing or establish new Current Research Information System (CRIS) for structural biology community. Data equivalent to information covered by CERIF standard are stored in multiple services (e.g. information projects, persons and organizational units are held by Instruct and offered via ARIA API) . West-Life services including Virtual Folder is prepared to provide data in the mentioned standard format.

3.6 Domain Specific Metadata

EUDAT service B2Share allows to share and annotate data using domain specific tags. Related service B2Note allows additionally tag dataset with additional tags not available among shared B2Share disciplines.

3.7 Future Work

Virtual Folder generates a provenance record which contains basic information, but misses the connection to 1) software, 2) software parameters, 3) original article and other information not mentioned above. This information can be manually added using the PROV-N editor, however, it is intended as a last resort for the data curator who may be able to put information in PROV-N notation standard. We will implement an automatic way for processing services to generate these missing records, including software parameters used and possibly published article or document describing the dataset or data file. Software can be identified by its home page. Specific version of the software can be identified by an URL to its endpoint, container description (e.g. Dockerfile, Vagrantfile) or template image (e.g. virtual machine image in EGI AppDB). Software parameters can be recorded from logs, or from forms filled by users within user interface. There could be designed a standard way to obtain parameters of computation, or aggregate specific parameter location for finite number of software usually used by structural biologists and automatically add such information into provenance record.

The provenance record is expected to provide information which should be sufficient to reproduce the computation from source data using the appropriate software tool and parameters, thus convert a specific provenance chain to a workflow.

A related standard is the Core Scientific Metadata Model (CSMD) - <http://icatproject-contrib.github.io/CSMD/>

The provenance record is part of metadata, therefore virtual folder and related service should be considered to provide more metadata in CERIF format. This will allow common research information systems (CRIS) to harvest public data and metadata automatically and make them discoverable and available for broader community.

Virtual Folder supports storing provenance records in ProvStore service within <https://openprovenance.org> which is dedicated to store any provenance documents conforming PROV-O standard. EUDAT recently released a B2Note service allowing custom annotation of any data stored within EUDAT storages, this might be considered as another form to store provenance or custom metadata.

References cited

<https://www.w3.org/2005/Incubator/prov/wiki/images/0/02/Provenance-XG-Overview.pdf>.

Core Scientific Metadata Model (CSMD) - <http://icatproject-contrib.github.io/CSMD/>

ProvStore service - <https://openprovenance.org>

User's guide <https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/users-guide/metadata>

Developer's guide <https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/developers-guide/metadata-and-api/dataset-metadata-and-api>