

Work Package 4, Milestone 4.3: Prototype access to selected datasets

Milestone report written by Tomas Kulhanek STFC

Many services rely on existing external data sets. The infrastructure will have to be capable of leveraging this external data upon user's request. This milestone 4.3 (M13) is part of Task 4.3 and is based on the review of the relevant datasets (reported as D4.4). Selected datasets are integrated into the WP6 Virtual Folder and further issues and thoughts are discussed. Some of the related work in progress is tracked within WP6 wiki pages <http://internal-wiki.west-life.eu/w/index.php?title=D6.2Metadata>, therefore, this report is a snapshot of the relevant information available there at the time of this milestone, i.e. September 2017.

Data sources

Within deliverable D4.4 these data sources with datasets were identified, for further details refer D4.4.

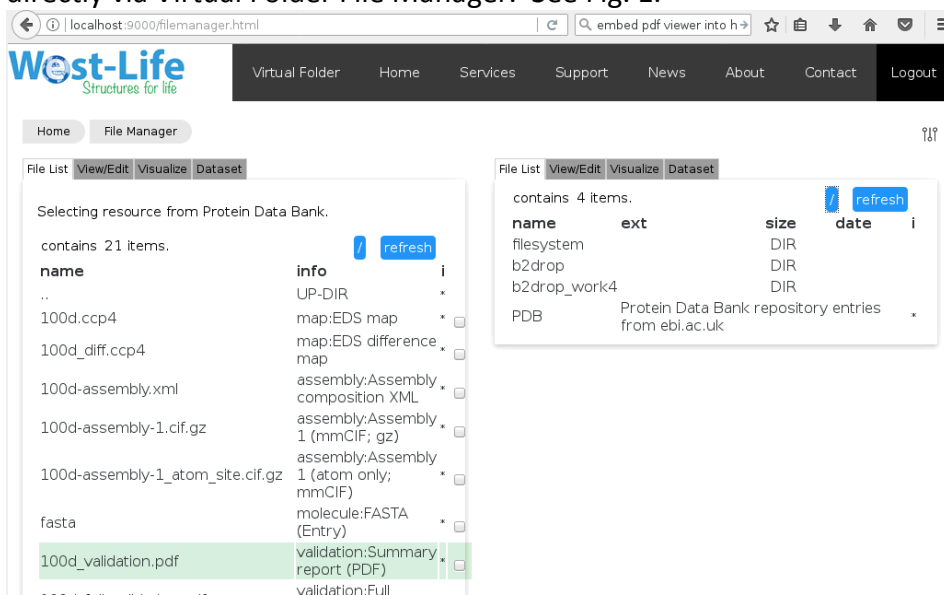
1. Experimental data at synchrotrons – implemented in partners STFC facility Diamond Light Source. Uses iCAT data management system (<https://icatproject.org/>).
2. Protein Data Bank – maintained by the project partner EMBL EBI. It offers RESTful API and web components to obtain metadata and data from PDB.
3. NMR Data at BioMagResBank (BMRB) – a European mirror maintained by project partner CIRMPP. It offers RESTful API to obtain NMR data related to BMRB entry. Other API offers HTML output.
4. PDB-REDO archive – maintained by project partner NKI. It contains PDB files with improved structures by PDB-REDO software.
5. Uniprot – Project partner EMBL EBI is member of Uniprot Consortium. Uniprot offers RESTful API and web component is provided.
6. Experimental data at EM centres – Project partner EMBL EBI offers EMPIAR repository, to access primary EM data (hundreds of megabytes to few terabytes of data in the form of movies) and allows to download this data using Aspera technology for fast network transfer.
7. Experimental data at other facilities – this is going to be addressed in deliverable D6.2 to provide exemplar installable repository with metadata and standard API to be easily integrated with other West-Life services
8. Data linked from publication – standard DOI link to refer to any above mentioned repositories or generic repositories like Zenodo (<https://zenodo.org>).
9. Other data repository initiatives – SBGrid repository (<https://data.sbgrid.org>), provides RESTful API to query datasets, some of them are related to published structural models.

Milestone M13

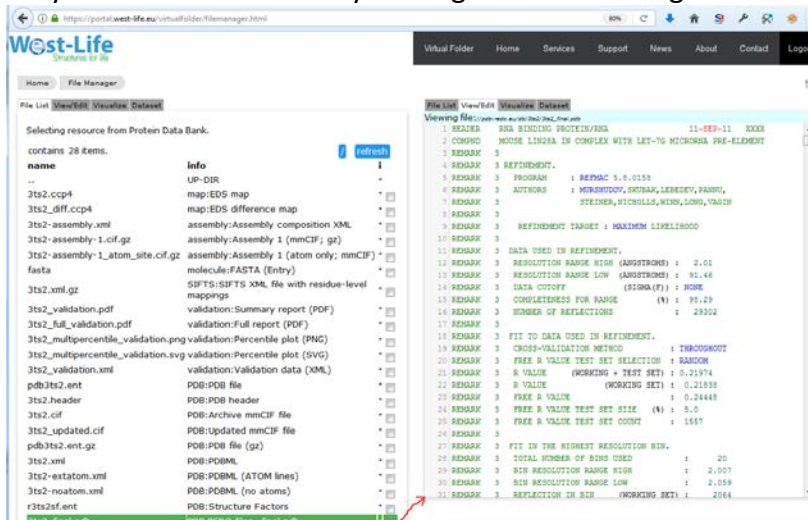
Access to selected datasets in data sources

Per the identified data sources the following were done towards integration to West-Life services:

- ad 1. Experimental data at synchrotrons – Integration, access to data via iCAT interface can be developed using RESTful API. There was considered to implement a WEBDAV interface towards iCAT data catalogue, thus, be able to be directly integrated into West-Life Virtual Folder. However, mapping between WEBDAV protocol and iCAT API was recognised as to be too complex.
- ad 2. Protein Data Bank – Prototype dataset access was made with autocomplete component and web components, see D5.5 for full report. Additionally, “PDB” virtual folder was added as a prototype access to PDB entries and all linked dataset files directly via Virtual Folder File Manager. See Fig. 1.



- ad 3. NMR Data at BioMagResBank (BMRB) - Not yet integrated to dataset prototype. Some of the API returns html output, thus in order to connect related datasets with relevant PDB id's, further proposal to output format might be addressed by BMRB.
- ad 4. PDB-REDO archive - “PDB” virtual folder contains links to relevant PDB-REDO files if they exist for the PDB entry. See fig. below showing source of improved PDB file



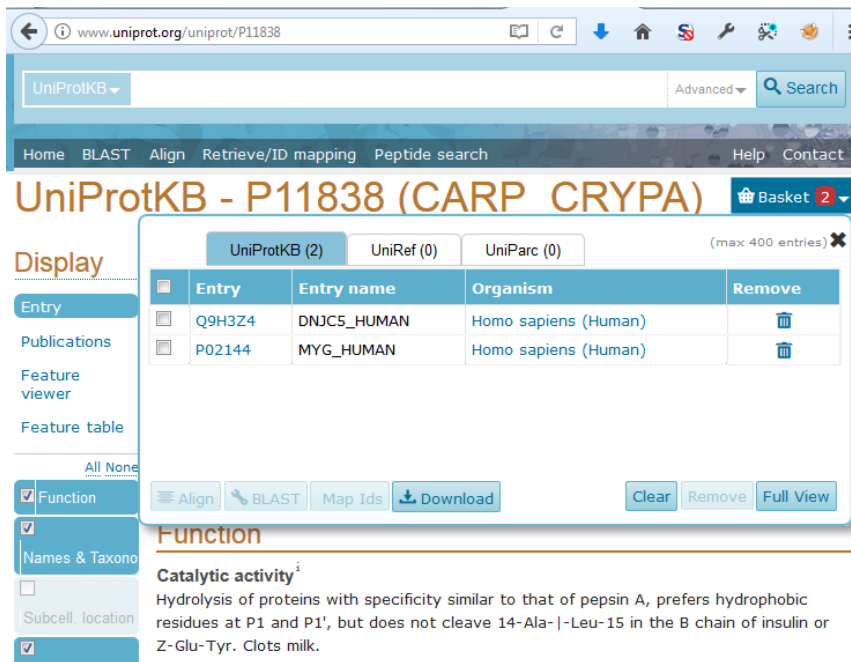
Milestone M13

ad 5. Uniprot - The uniprot web component is integrated into Dataset prototype. See D5.5 for full report and figure below.

The screenshot shows a web browser window with the URL <https://portal.west-life.eu/virtualfolder/filemanager.html>. The page title is "West-Life Structures for life". The navigation menu includes "Home" and "File Manager". The main content area has tabs for "File List", "View/Edit", "Visualize", and "Dataset". A text box contains "dataset-2017.5.12" and a label "PDB or related item to add:" with a text input field containing "P11838". Below this is a "PDB Prints" dropdown menu. A list of items shows "P11838" with a close icon and a link icon, labeled "recognized as UniProt entry" and "UniProt Link P11838". A "PDB UniProt Viewer" window is open, displaying "Endothiapepsin" (P11838) from *CARP_CRYPA* *Cryphonectria parasitica* EAPA. It shows a sequence map with 525 chains in 525 PDB entries and a list of UniProt accessions: 5hct, 1cew, 1gvu, 1gvt, 4y5l, and 1gvw. Each accession has icons for X-RAY, 3D structure, and download.

ad 6. Additional links could be obtained via appropriate RESTful API and added into dataset file list as of points 2 and 4. However, Uniprot web portal offers to select dataset files and collect them into virtual "basket" which can be downloaded later in one file archive. This might be enhanced with Virtual Folder's Upload Dir Picker component – to select directory from West-Life Folder which the data will be downloaded into by the server without involvement of user computer. See current basket in screenshot below.

Milestone M13



UniProtKB - P11838 (CARP CRYPA)

Entry	Entry name	Organism	Remove
Q9H3Z4	DNJC5_HUMAN	Homo sapiens (Human)	
P02144	MYG_HUMAN	Homo sapiens (Human)	

Function

Catalytic activityⁱ
 Hydrolysis of proteins with specificity similar to that of pepsin A, prefers hydrophobic residues at P1 and P1', but does not cleave 14-Ala-|-Leu-15 in the B chain of insulin or Z-Glu-Tyr. Clots milk.

ad 7.

ad 8.

ad 9. Experimental data at EM centres - EMPIAR data access is not yet integrated in dataset. EMDB pages now link to the 3dbionotes analysis service provided by partner CSIC. Additional links to the EMDB database and to the 3dbionotes analysis service could be included into relevant West-Life Virtual Folder part. Further analysis about this idea should be done.

ad 10. Experimental data at other facilities – Currently the DSpace and Dataverse software are considered as a base repository system for small facilities. Both software implements relevant metadata standards and RESTful API to discover datasets published into it.

ad 11. Data linked from publication – no implementation yet. The relevant data from publication can be discovered via appropriate search query and added into list of files (points ad 2. and ad 4.)

ad 12. Other data repository initiatives – the documented RESTful API of SBGRID seems to primary link to datasets and discovery of e.g. related datasets to PDB entries seems not to be supported yet or returns HTML output only. Links to published datasets could be added to dataset files (points ad 2. and ad 4.).

Milestone M13

Conclusion

These main types of data access is available

- West-Life Virtual Folder web application provides link to user – see ad 2., 4. E.g. as autocomplete search currently available as "dataset" tab for PDB entries or a "special folder" allowing to browse repository and upload it to Virtual Folder.
- Third party web application provides a feature like basket – which can be downloaded in batch. A button “Download to West-life Virtual Folder” can be introduced to initialize file transfer between repository and West-Life Virtual Folder instance which can be different computer on network backbone. See ad 5.
- Third party portals usually provides domain specific UI and rich application – West-Life can leverage this by providing very specific links to the relevant web application.
- Third party portal provides web component which can be integrated into any other UI – e.g. pdb components used in points 2. and 4. West-Life UI can integrate these components into relevant UI context.