Deliverable D4.4

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	**675858**	
Deliverable title:	**Overview of external datasets, strategy of access methods, and implications on the portal architecture**	
WP No.	4	
Lead Beneficiary:	4: Masarykova University	
WP Title	Operation	
Contractual delivery date:	30 April 2017	
Actual delivery date:	28 April 2017	
WP leader:	Ales Krenek	MU
Contributing partners:	STFC	

Deliverable written by Chris Morris STFC, Tomas Kulhanek STFC

Contents

1 Executive summary

Structural biologists are undertaking more challenging research projects, requiring a greater variety of techniques, and so visiting wider range of research facilities. Tracking, processing, and sharing their data is getting increasingly complex. Web services provided or coordinated by West-Life must be able to access these data, and also to record metadata about their use.
This task reviewed the relevant datasets to be made available, and defines architecture and appropriate interfaces to access them. It built on the metadata services to be offered in WP6. Together with T4.4 security issues (authorization and user identity delegation in particular) were addressed.
Wherever feasible, access to datasets will be a part of the integrated “Virtual Folder”. However, some cases are reported where the available interfaces are not mappable to a directory tree. In these cases, web components will be developed to make it easier for service providers to integrate access to these data sources. This particularly applies to records from public databases.

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**		x
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**		x
4	**Foster best practices, collaboration and training of end users**	x	

3 Detailed report on the deliverable

3.1 Data Sources

Experimental data at synchrotrons

Beamlines for X-ray crystallography are mature infrastructures with sophisticated data management and processing that has been built up over previous projects (e.g. http://pan-data.eu/). Data collected at synchrotrons and neutron sources are stored in facility systems, which in the case of ESRF, DLS, and ISIS use the open source meta-data management system iCAT [https://code.google.com/archive/p/icatproject/]. The STFC will develop an interface to iCAT to make it part of the Virtual Folder, so that these datasets can be imported by partner portals. Synchrotrons also store additional metadata about the user visit, and the experiments and data processing undertaken (e.g. ISpyB at Diamond).

Protein Data Bank

This is the public repository for macromolecular structures, operating since 1971. The world-wide PDB [https://www.wwpdb.org/] includes PDB-Europe [https://www.ebi.ac.uk/pdbe/node/1], managed by the West-Life partner EMBL EBI. The wwPDB's initiatives include data standards, validation task forces, and a common system for deposition and validation.

Partner EBI is developing a RESTful API for searching and obtaining data from the PDB as part of the work in WP6. With support from STFC, it is developing web components to make this easily reusable in web portals offering structural services.

NMR data at the BioMagResBank (BMRB)

BMRB (http://www.bmrb.wisc.edu/bmrb/) collects, annotates, archives, and disseminates (worldwide in the public domain) the spectral and quantitative data derived from NMR spectroscopic investigations of biological macromolecules and metabolites. BMRB is a partner of the worldwide PDB. Partner CIRMMP is hosting the only mirror in Europe of the BMRB (http://bmrb.cerm.unifi.it/).
The recently developed NMR Exchange Format (NEF, doi:10.1038/nsmb.3041) is a standard format for NMR-based restraints and associated data.

PDB-REDO archive

This archive [http://www.cmbi.ru.nl/pdb_redo/] contains downloadable PDB files representing improved structures created by the PDB-REDO service, maintained by the West-Life partner NKI. Compared to the structure files originally deposited in the PDB, those in PDB_REDO have been re-built and re-refined using the latest protocols and software, and are often significantly better by current quality standards.

Many structural biology web services accept PDB files as an input. We will make it easier to choose a file from PDB-REDO by providing a web component that the service provider can add to the submission form.

Searches using the PDB API mentioned above can also find structures in PDB-REDO, because the same accession codes are used. West-Life will provide web components to support this.

UniProt

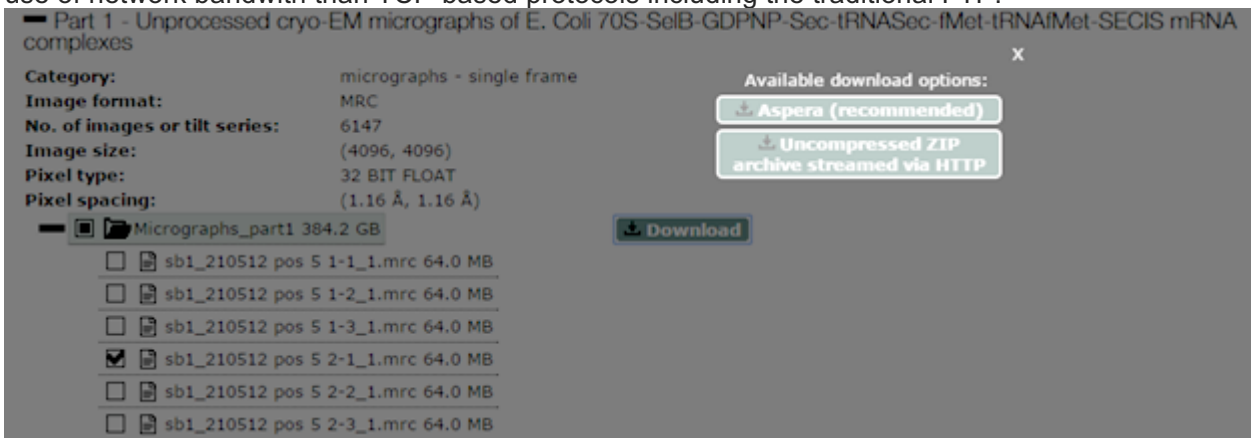
Universal Protein Resource [http://www.uniprot.org/] is a catalog of information on specific proteins. Its knowledge base database consists of nearly $86x10^6$ fragments, whose protein existence comes from the structural information of proteins (0.16%) which is linked to Protein Data Bank, proteins whose evidence about existence comes from transcript level (1.34%), inferred from homology (23%) and predicted (74%).

Partner EMBL-EBI collaborates in the Uniprot consortium. A RESTful API is available and a web component PDB Uniprot Viewer is developed to provide links between Uniprot entries and related PDB entries and to make this easily reusable in web portals too, using the new service Structure Integration with Function, Taxonomy and Sequence [http:// pdbe.org/sifts].

Experimental data at EM centres

New detectors have made Cryo Electron Microscopy more powerful, with the first 1.8 Angstrom structure in 2016, leading to an expansion of the use of this technique and the number of centres offering it.

In a visit to an experimental infrastructure site lasting one or two days, scientists acquire one to two terabytes of data in the form of movies. The next generation of detectors is likely to acquire data faster than it can be written to disk, forcing the use of SSD drives or tapes. The Electron Microscopy Public Image Archive (EMPIAR) provided by the EMBL-EBI partner offers this primary EM data for download, when a user chooses to deposit this data. The sizes vary from hundreds of megabytes to a few terabytes. The EBI therefore recommends download by Aspera, which makes more efficient use of network bandwith than TCP-based protocols including the traditional FTP.



Scientists usually take the data home from the microscope facility on a USB drive. The first step of processing reduces it by a factor of ten to a set of averaged video frames, usually referred to as micrographs, and a Contrast Transfer Function. Then guided computations produce particles, classes, and eventually a 3D electrostatic potential map. Currently, this stage can take weeks or months, but rapid advances in software are reducing this, such that in the near future it may be less than a day for

favourable cases. The 3D map must be deposited in EMDB. Furthermore, in those cases where map resolution is high enough for atomic modelling, maps are refined to produce atomic coordinates, which are deposited in the PDB linked to the map deposition at EMDB. EMDB pages now link to the 3dbionotes analysis service provided by partner CSIC.

Scientists adopting cryo EM methods face significant challenges in sample preparation and computational skills, as well as in provisioning compute resource capable of handling these movies. Some of these barriers would be reduced if facilities could offer on-site processing facilities, so that processing can occur close to the data, at least for the steps that require access to the movies, that is to say the initial data reduction leading from the movies to the micrographs.

In order to ensure that the dataset is still available at the latter stage, storage is needed that is adequate for at least 6 months of data acquisition, of the order of 100TB. An archive of this size is most cheaply provided as a tape store. The eBIC centre in the UK has access to the Atlas data store, but most European centres do not currently offer archives.

West-Life will provide access to metadata from such facility datastores and EMPIAR. We will also seek to publicise these examples of best practice, so that in future acquisitions of electron microscopes proper thought is given to the resulting needs for complementary information technology.

Experimental data at other facilities

Instruct offers access to 13 techniques for structural analysis, 9 biophysical techniques, and 15 techniques for sample preparation, at 24 centres [https://www.structuralbiology.eu/platform-catalogue]. So, a very wide range of data is potentially relevant to the provenance and results of a structural project.

Instruct's Data Management Policy says *"storage of data is the responsibility of the User to whom it belongs ... Instruct Centres are not required to take responsibility for storing data beyond the immediate acquisition visit or the time taken for post experimental analysis if the latter is also provided by the Centre. However, Instruct Centres aspire to offer an archive to store data, especially in cases where the data volume makes this more practical than transferring the data ..."*

In contrast to the image acquisition for crystallography and EM described above, the primary data for most of these techniques are not large (although notably for NMR it becomes larger during the initial processing stages). Users often take their data home on a USB stick or through download from an ftp site.

It would be better if the facilities could provide a data management service, compatible with the aspiration stated in Instruct's Data Management Policy for Centres. In order to reduce the barriers to doing this, D6.2 will be an installable repository for the use of Instruct centres.

However, it is also desirable to be able to provide globally unique IDs for datasets on mountable media. This is discussed below.

Data linked from publications

Increasingly, journals require that experimental data is open and that a paper links to it with a Digital Object Identifier (doi). These data may be in one of the discipline specific repositories mentioned here, a university repository, or a general repository like Zenodo.

Increasing numbers of these repositories conform to the Datacite metadata standard [https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf] which West-Life will therefore support.

Other Data Repository initiatives

A recent initiative supported by the National Institute of Health is the SBGrid repository in the US (http://data.sbgrid.org). It has been established as a diffraction data publication and dissemination system to preserve primary experimental data sets that support scientific publications, but also accepts modelling data (Sliz et al. Data publication with the structural biology data grid supports live analysis. Nature Communications, 7: 10882, 1-12 (2016)). Data sets are accessible to researchers through a community driven data grid, which facilitates global data access. For example, several datasets from the Utrecht partner related to HADDOCK can be found on the SBGrid data repository: https://data.sbgrid.org/search/?q=Bonvin

3.2 Strategies for data access

In the light of the details above, several different IT strategies are needed to support access to and sharing of structural datasets and the recording of appropriate metadata.

File system access

Structural biologists have been using computers since Dorothy Crowfoot Hodgkin's work on penicillin in the 1940s. Many of the programs they use are still designed for command line use, and expect to read their input files directly from a local file system.

We can relieve users of the need to copy the file by using the CVMFS file system with the DIRAC user interface [https://twiki.cern.ch/twiki/bin/view/CLIC/DiracForUsers] to mount their data in the West-Life Virtual Folder, or a WebDAV interface. Work is needed to make files in the facility repository iCAT available this way.

Demountable file systems

As discussed above, files are often carried on USB devices, either because they are small and little attention has been given to data management services, or because they are too large to transport any other way.

For large data, transporting a data storage medium scales well as an alternative to sending the data over a network. However, tape may do better than NAND flash memory. The industry roadmap for tape storage projects capacity is increasing by 41%pa [http://www.lto.org/wp-content/uploads/2014/06/2015-Technical-Roadmap.pdf]. CISCO predicts a compound annual growth rate of 22% for the volume of data transferred over the internet [http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html]. This makes dispatching a tape not an obsolete method, but more and more competitive over time.

We will add support in the West-Life Virtual Folder for demountable devices, with a persistent GUID to identify the files.

API Access

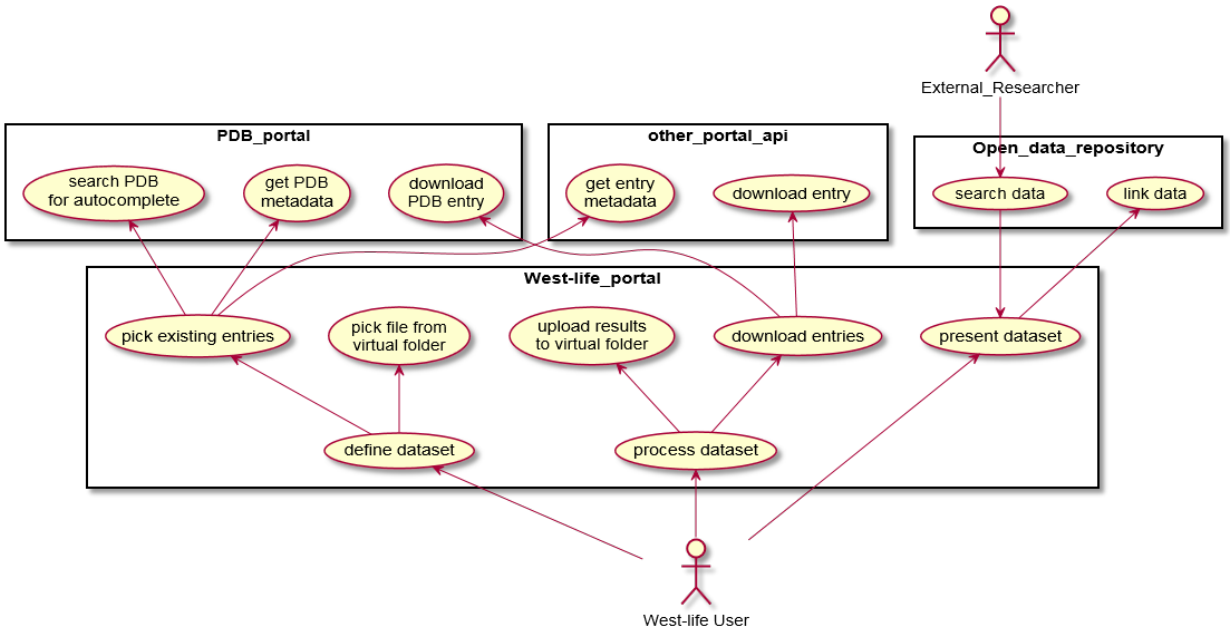
Some of the services mentioned above do not provide interfaces suitable for file system mounts, but do provide an API for downloading the file. West-Life will provide code to do this, allowing the use of existing programs without modification.

This applies to files from databases like the PDB, and to files linked from publications.

3.3 Implication on portal architecture

The metadata of a dataset provides information about the dataset, its provenance and its ownership and relation to projects, publication, other datasets etc. The use case diagram below shows relation

to other use cases of the West-Life portal or partner (PDB) portal which are already implemented or expected to be available. Additionally, the dataset should be searchable by an open data repository.



An example of an open data repository network in Europe is Zenodo [https://zenodo.org/]. In order to interlink and register a west-life portal and its public dataset into open access infrastructure database, the following guidelines must be met. CERIF standard is recommended as the format of metadata of datasets, see the CERIF entity cfResultProduct [https://zenodo.org/record/17065]. Datasets are linked with publications, with funded projects, with persons and organisations, and with equipment. Additionally, REST api must be implemented in order to return the metadata of dataset entries at */resultproducts* endpoint as defined by the standard.

XML format is mandatory by default, JSON format is not mandatory. As XML is generated by the current framework of metadataservice of VF (ServiceStack), it might be investigated whether the generated format will be automatically compliant with the XML Schema defined by CERIF standard based on dataset (cfResultProduct) structure, otherwise customization should be made.

The current implementation of metadataservice covers defining the dataset utilizing some of the pdb-component library. The process dataset is partly covered by tools already integrated into west-life portal (tools in custom VM and ongoing integration with WENMR portal). The use case “present dataset” is not yet designed nor implemented. Therefore, the dataset should be presented as:

1. Metadata structure - which can be browsed via RESTful API to obtain details, or links to the data sources mentioned above. The API and format returned should follow the CERIF standard
2. Directory in virtual folder, therefore accessible via WEBDAV using a generated link - each entry can be a subdirectory, file or link. Proper mapping to WEBDAV should be implemented e.g. by downloading and exposing entries as files to the current WEBDAV interface.
3. File - ZIP file with the directory structure as defined above. For bigger datasets, this might not be practical and can be limited to an archive of small entries.

References cited

https://www.structuralbiology.eu/platform-catalogue

http://www.lto.org/wp-content/uploads/2014/06/2015-Technical-Roadmap.pdf

https://www.openaire.eu/

https://zenodo.org/record/17065

https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf

Appendix 1: Summary of data sources

Protein Data Bank	
Where	In an online database
What	.pdb/.mmcif plus .mtz, tens of Kb
How	Make URL containing accession code or use search API
Possible output	Usually revised .pbd/.mmcif, e.g. after Molecular Replacement
Where	In an online database
What	.pdb/.mmcif plus .mtz, tens of Kb
How	Make URL containing accession code or use search API
Possible output	Usually revised .pbd/.mmcif, e.g. after Molecular Replacement

Experimental data at synchrotrons	
Where	Usually in iCAT repository
What	Diffraction images, 100s of 1-10 Mb files
How	iCAT UI to request staging from tape
Possible output	Merged reflections, a few Mb.
Experimental data at EM centres	
Where	eBIC, CSIC, etc
What	Several 100s of movies per day, 1.5 GB each (800 movies/day = 1.2TB/day). Microscopes could render 1 movie per minute, only during acquisition time. Some time is invested in setting/configuring, loading or screening areas.
How	West-Life will provide access to metadata on these large datasets. We will also publish on the IT requirements for future microscopes.
Possible output	Particle images, a few Mb

Experimental data at other facilities	
Where	Experimental facility e.g. CERM
What	Various, Kb to 1Gb
How	Plug in a USB drive
Possible output	Usually reduced data. For NMR: spectra, Gb.
Data linked from publications	
Where	Zenodo, B2SHARE, university repository, and as above
What	Any
How	Resolve doi
Possible output	Any