Deliverable D7.9

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	675858	
Deliverable title:	Report on existing metadata standards, and proposals for new vocabularies	
WP No.	7	
Lead Beneficiary:	1: STFC	
WP Title	Joint Research Activity	
Contractual delivery date:	30 April 2018	
Actual delivery date:	30 April 2018	
WP leader:	Jose Maria Carazo	CSIC
Contributing partners:	STFC, EMBL-EBI, UU	

Deliverable written by Chris Morris

Contents

Executive summary

Structural biology has a long tradition of open data, notably through depositing structures in the Protein Data Bank. Nevertheless there are challenges to transition to full compliance with the FAIR principles. As reported in M6.3:

- Current workflows in Structural Biology may not be properly described in an unambiguous manner due to the lack of appropriate metadata standards specifying them (M6.3 report).
- There is no agreed ontology for the primary data processing, either at the level of integrated studies combining different technologies or even at the single technique level (M6.3 report).
- Metadata about experimental conditions are often incomplete
- As a result of this, the chain of custody from sample to publication is often broken at several points

This report describes progress on these issues:

- Extensions to the mmCIF data standard and increasing use of it,
- … particularly for hybrid and integrative models,
- Recording provenance
- Linking datasets to research projects
- Adoption of CWL for describing workflows
- Metadata support in West-Life software

The pressing needs that we found were not for novel metadata vocabularies, but for take up of existing ones. This report therefore plans practical steps to deliver improvements during remaining months of West-Life, and steps to enable future culture changes and implementation.

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 Adoption of mmCIF

Macromolecular structures have been represented in PDB-X files, with extension .pdb. However this format has limitations that make it unsuitable for larger structures, so in 2007 a task force recommended a transition to the mmCIF format for deposition and processing. Regretably ten years later this transition was by no means complete, despite leadership by partner EMBL-EBI.

West-Life partners have made significant progress towards use of mmCIF:

- DipCheck now accepts mmCIF input
- ARP/WARP accepts mmCIF for ligands but not proteins (EMBL-Hamburg)
- The PDB-REDO databank now stores mmCIF files
- The PDB_REDO service reads and writes mmCIF
- Version 8 of PDB-REDO will use mmCIF internally (NKI).
- The new HADDOCK portal accepts mmCIF input
- The DISVIS portal accepts mmCIF input
- 3DBionotes reads mmCIF

It is also worth noting that work in progress in CCP4 will lower the energy barrier to use of mmCIF for developers by integrating support into a standard library.

PDB-Dev (https://pdb-dev.wwpdb.org/) is a prototype deposition and archiving system for structural models obtained through integrative/hybrid (I/H) methods, as reported in D3.5. Among the members are Sameer Velankar, from partner EMBL-EBI, and Alexandre Bonvin from partner UU is a member of the wwPDB Integrative Methods Task Force. Because of the size of the models and need to record novel forms of provenance, these models are expressed in an extended dialect of mmCIF.

The first HADDOCK model based on integrative modelling has been deposited in the PDB-Dev database (https://pdb-dev.wwpdb.org/) (Accession code: PDBDEV_00000014) in February 2018 by partner UU. This model is a refined structure of the methyltransferase KsgA bound to the 16S ribosomal RNA subunit in E. coli. It was the result of a new HADDOCK protocol integrating cryo-EM maps information to drive the docking together with mutagenesis and DNA footprinting data to identify key residues for the interaction. In order to achieve this, Mikael Trellet visited RCSB to agree extensions to represent distance restraints.

3.2 Provenance

Each publication in structural biology is derived from processing steps, using data from a structural experiment, using a sample 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.

The W3C recommendation PROV-O [PROV-O: The PROV Ontology] is divided into several layers. The “starting point” supports the following concepts:

- *Entity*
- *Activity*
- *Agent*

These can be used to identify samples and datasets; experiments and processing steps, and scientists, respectively.

It also defines these relationships:

- *Generated by*
- *Derived from*

These are sufficient to describe the relationships between these entities.

The Expanded and Qualified terms in PROV-O allow recording more data, including the dates of the activities.

We concluded that PROV-O is sufficient to record the derivation of structural results. We therefore plan that D6.2 will implement support for PROV-O in the West-Life infrastructure, both the Virtual Folder and the Repository.

D6.2 will also supply a Python API for creating provenance records automatically as a service is executed. There is substantial evidence that scientists do not invest time in creating metadata. The automatic acquisition of metadata as they perform their research is now possible as a result of the growing use of web services rather than installed software.

We made a successful application to become an EUDAT Data Pilot. In addition Instruct has requested recognition as an EUDAT community for B2SHARE, with the core PROV-O concepts added as community metadata terms:

wasDerivedFrom	Records another dataset that was involved in the creation of this one.
hadPrimarySource	The experimental data that was the basis for the derivation of this dataset
wasRevisionOf	A revision is a derivation that revises an entity into a revised version.
alternateOf	Two alternate entities present aspects of the same thing. These aspects may be the same or different, and the alternate entities may or may not overlap in time.
wasGeneratedBy	Records a processing activity that created this dataset
wasInvalidatedBy	Records an activity that made this dataset obsolete
wasAttributedTo	Records the scientist or other agent which is the author of this dataset

3.3 Projects

Scientists naturally associate their research datasets with the research projects they are part of – not with the date of visit when they were recorded or other possible categorization.

The Common European Research Information Format [CERIF] enables recording research projects, investigators, publications, services, and datasets. OpenAIRE uses this standard at the time of ingest. We will therefore make the metadata imported from ARIA and saved in the West-Life Repository compliant with CERIF.

This will point in the direction of federated search of data repositories, using search terms that are scientifically meaningful.

3.4 Workflows

The well-established structural biology codes like RefMac perform complex analyses consisting of many steps, which are configurable by the choice of keywords. They therefore contain an internal representation of workflows. However, these implementations predate the development of common workflow managers and metadata standards for workflows, and retrofitting them is infeasible.

The situation is different in cryo-Electron Microscopy. Because of the resolution revolution, this field is growing rapidly with much software development. Therefore partner CSIC is engaged in an EOSC Pilot Project 'CryoEM workflows: Linking distributed data and data analysis resources as workflows in Structural Biology with cryo Electron Microscopy: Interoperability and reuse'.

There are several workflow engines in use for scientific processing, including Galaxy and Taverna, and in industry Pipeline Pilot. These were entirely separate developments. The Common Workflow Language (CWL) has now defined a standard for exchange of computational workflows. We encourage new structural biology services to use this to express and execute workflows.

Prior to information processing, structural biology workflows include laboratory work. In order to make this more traceable, we have started discussions with OpenAIRE about the desirability of making experimental protocols citeable.

3.5 Metadata in the West-Life Virtual Folder

Storage provider and *file*/*directory* are the basic entities in the virtual folder. *Dataset* is another optional entity, used to store the location of data. The description of metadata about entities and how to access them are documented at https://h2020-westlife-eu.gitbook.io/virtual-folder-docs/virtual-folder/developers-guide/metadata-and-api . The automatic metadata of all operation and entities are generated by the framework at https://portal.west-life.eu/metadataservice/metadata, as documented at https://docs.servicestack.net/metadata-page.

For discoverability, the virtual folder additionally implements some Well-Known URI e.g. host-meta specified by RFC 6415 (https://tools.ietf.org/html/rfc6415) at https://portal.west-life.eu/.well-known/host-meta.

Partner INFN is now working on an extension to integrate the Virtual Folder with OneData.

References cited

Sali et al, Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop, Structure. 2015 Jul 7; 23(7): 1156–1167.

Lebo et al, *PROV-O: The PROV Ontology*, 30 April 2013, https://www.w3.org/TR/prov-o/

CERIF https://www.eurocris.org/cerif/main-features-cerif

CWL Peter Amstutz, et al. (2016): **Common Workflow Language, v1.0**. Specification, *Common Workflow Language working group*. https://w3id.org/cwl/v1.0/doi:10.6084/m9.figshare.3115156.v2