Deliverable D6.1

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	**675858**	
Deliverable title:	D6.1 Virtual Folder	
WP No.	6	
Lead Beneficiary:	1: STFC	
WP Title	Data Management	
Contractual delivery date:	31 October 2016	
Actual delivery date:	31 October 2016	
WP leader:	Tomas Kulhanek	STFC
Contributing partners:	N/A	

Deliverable written by Tomas Kulhanek

Contents

1 Executive summary

We have built on previous work by WeNMR and others to create a virtual folder view of scattered data. It provides a consistent view of the files for a research project, regardless of the experimental facility in which they were obtained and regardless of their current location. This view is available through a web interface, and also as a mounted file system for access by programs which the user runs.

This work uses the B2DROP service provided by EUDAT. In order to facilitate use of EUDAT services by structural biologists, we worked with ARIA and EUDAT to ensure that Instruct userids are accepted by EUDAT's authentication service B2ACCESS [https://www.structuralbiology.eu/update/news/publish-data-with-b2share/]. Soon, these credentials will also be accepted by B2DROP – at the time of writing, they are already accepted by B2SHARE.

The output from this work is available as a Virtual Machine [VM] suitable for running locally and also for use on EGI resources.

This is be the first such effort to also address data management for the growing field of single particle electron microscopy, by including the SCIPION suite in the VM.

This tool has been registered in the ELIXIR Tools and Data Services Registry.

CCP4 has begun development of a cloud service for solving crystallographic structures. We have built a version of the VM that includes CCP4 code. This version will be available on the CCP4 web site. Because the CCP4 license does not permit redistribution, West-Life cannot publish this version of the VM ourselves. This licensing policy is a part of CCP4’s successful strategy for sustainability.

In order to facilitate reuse of the Virtual Folder as part of larger scientific appliances, we took a “devops” approach, in which the installation process is under configuration management as rigorously as the software components are.

The sources are https://github.com/h2020-westlife-eu/west-life-wp6.

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 Use Cases

A typical scenario follows the data life cycle as reported in Deliverable D3.1:

1) A structural biologist has collected a large dataset at a central experimental facility, which currently resides in data storage provided by the facility or by her home institution or by an infrastructure provider (section 1 of D3.1: Creating Data). Data reduction has been performed at the facility, and the output stored with the raw data (section 2 of D3.1: Processing Data: Data Reduction).Additionaly she has collected a large dataset at another experimental facility, which currently resides in another data storage.

2) She has an initial structural model and some additional restraint information, and wishes to refine the model against the newly acquired experimental data (section 3 of D3.1: Analysing Data: Structure Determination and Interpretation). Thus she uses the West-Life portal to launch a selected computational tool, and access both the data storage where the new experimental data resides.

3) After refinement, the resulting model is stored locally or back in the remote data storage. If the model satisfies quality control, then the experimental data, the resulting model, and a record of the refinement are submitted to the PDB database (section 4 of D3.1: Preserving Data and Giving Access to Data).

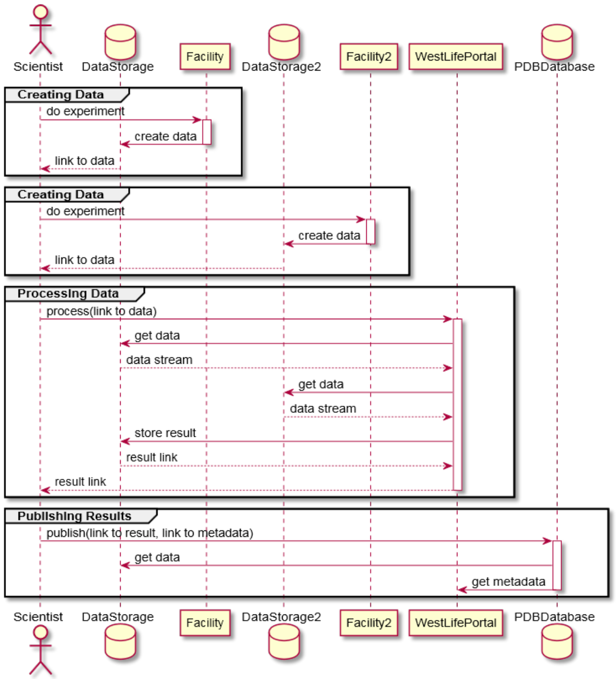


Figure 1: Scenario, how West-life portal is involved in various life-cycle stages of scientific data.

The goal of the Virtual Folder (VF) (part of WP6 – data management) is to integrate existing external and internal data repositories into the West-Life portal and give access via single interface to be used by other services (see Figure 2). In the above scenario, the VF coordinates access to the experimental facility data, the prior knowledge (initial model and restraint information), the results of further processing (refinement of the model), and deposition to a data repository such as the PDB. Integration of data storage in turn supports the use of computational tools which operate on that data, and also supports searching and linking of external data sources.

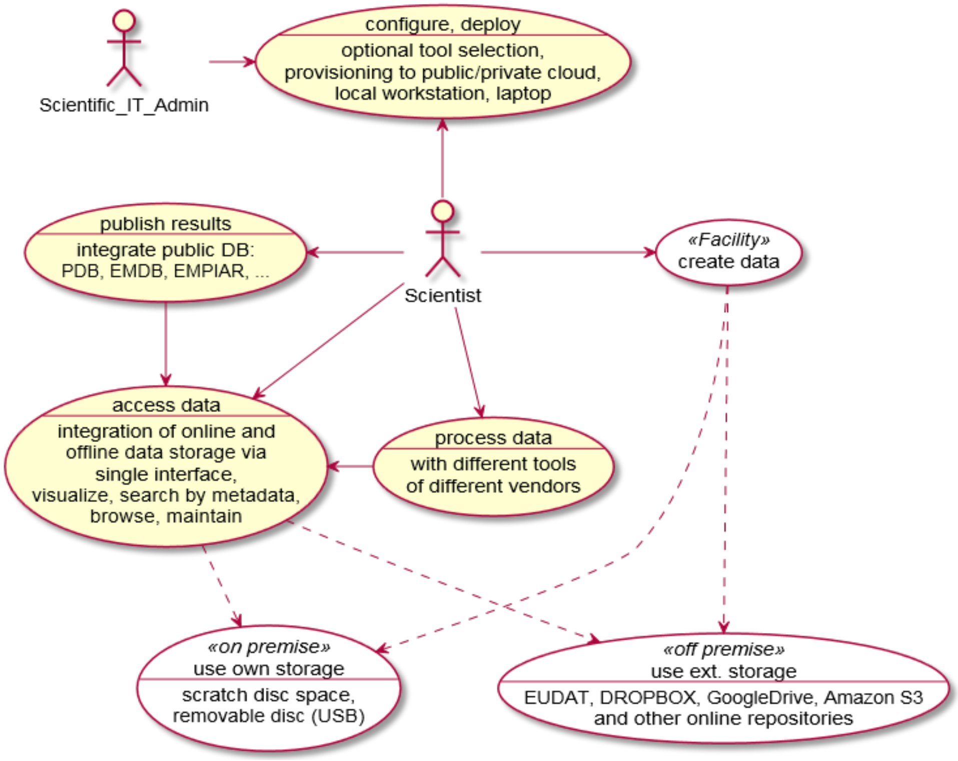


Figure 2: Use Cases addressed by Virtual Folder in yellow and related use cases in white

The Virtual Folder can be deployed in two ways: as a single-user virtual machine (VM), or as a part of the West-life Portal. In the first case, the user has full control of the VM and determines how it is accessed, and which data sources are included. In the second case, another user (labelled Scientific_IT_Admin in Figure 2) is responsible for configuring and deploying the VF as part of a larger e-infrastructure.

3.2 Single user deployment

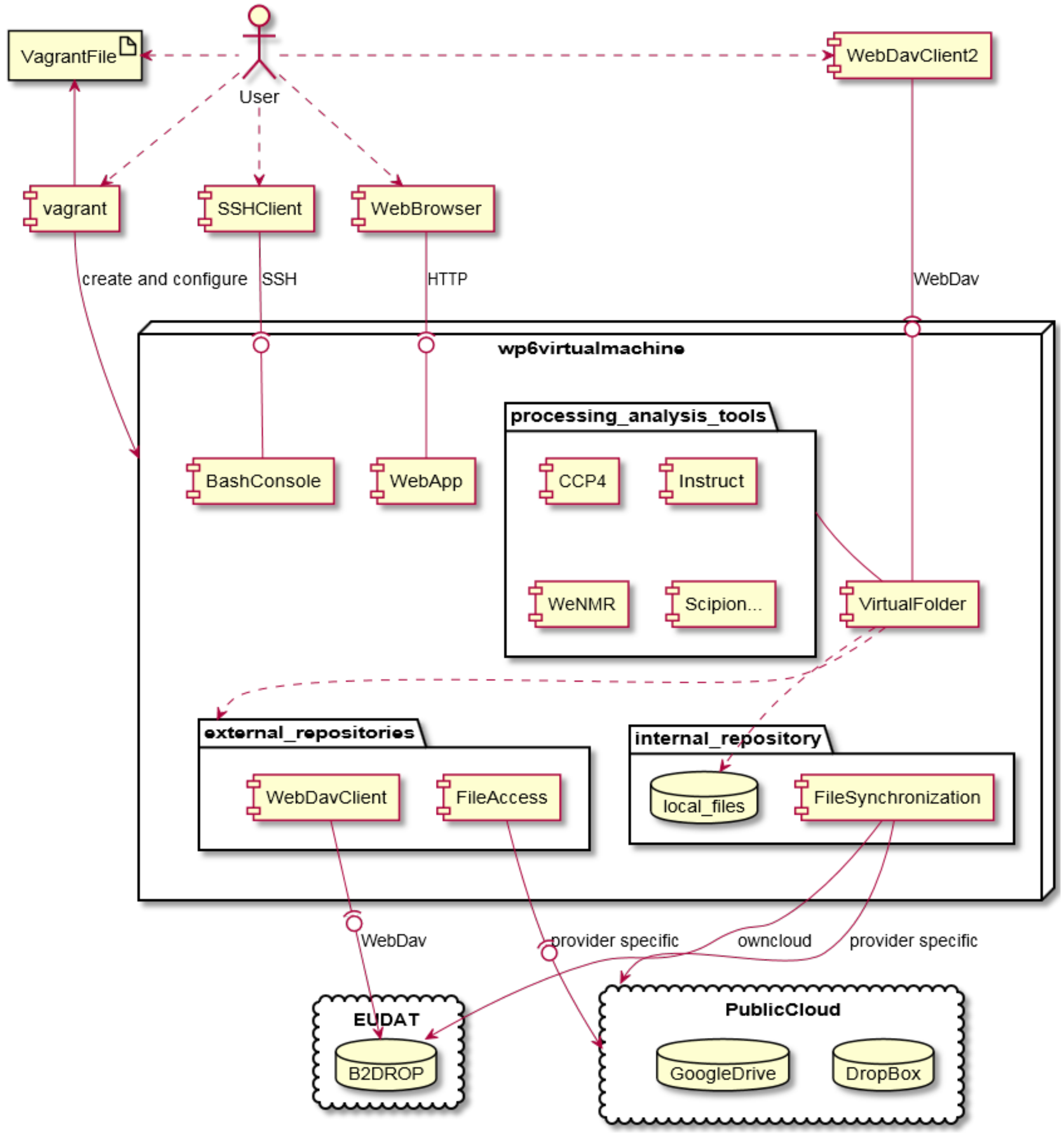


Figure 3: Components of VF in single user deployment

Downloading and running a single VM provides a stand-alone solution for structural biologists. Additionally this scenario is used for development and testing the system. In the current implementation, the user downloads a package containing the necessary source code, and the vagrant tool [4] together with VirtualBox are used to build and configure the VM. The VM runs a

Linux OS, contains pre-installed scientific software, and has pre-configured links to relevant data sources.

Once the VM is built and running, it can be accessed in a number of ways. As with other VMs, it can be accessed interactively using VirtualBox. Processing and analysis can be done within the VM using the pre-installed software, with data on the host OS accessed using Guest Additions in the usual way. Alternatively, a scientist can work on the host machine (or on a remote machine, if port forwarding is used) and ssh to the VM to access specific programs and data. In this mode, the WP6 VM is similar to other software projects that use virtual machines to distribute ready-to-use software (e.g. BioLinux http://environmentalomics.org/bio-linux/), with the main difference being the pre-configuration of relevant external data repositories for structural biology.

On the other hand, the VM can be treated as a black box, providing a set of services. A web interface is made available to browsers running on the host OS, which provides access to the main services. These include a folder view of all the connected data repositories, web interfaces to pre-installed scientific software, and a tool for connecting to new data repositories. The connected data repositories are also exported using the WebDAV protocol. These can be mapped to a drive on the host OS, and hence used as if they contain local data. In this mode, the scientist continues to work on the host machine, and the WP6 VM acts as a helper utility for accessing data repositories and web services.

Irrespective of the mode of operation, the virtual machine connects to data storage services (e.g. EUDAT's B2DROP) and can use the same mechanism to connect to other repositories. The connection information is kept including authentication tokens. The virtual machine can be run on the user's own hardware. For cases when more computational power is required, it is also suitable for deployment on scientific cloud resources, e.g. EGI FedCloud.

3.3 Westlife Portal Deployment

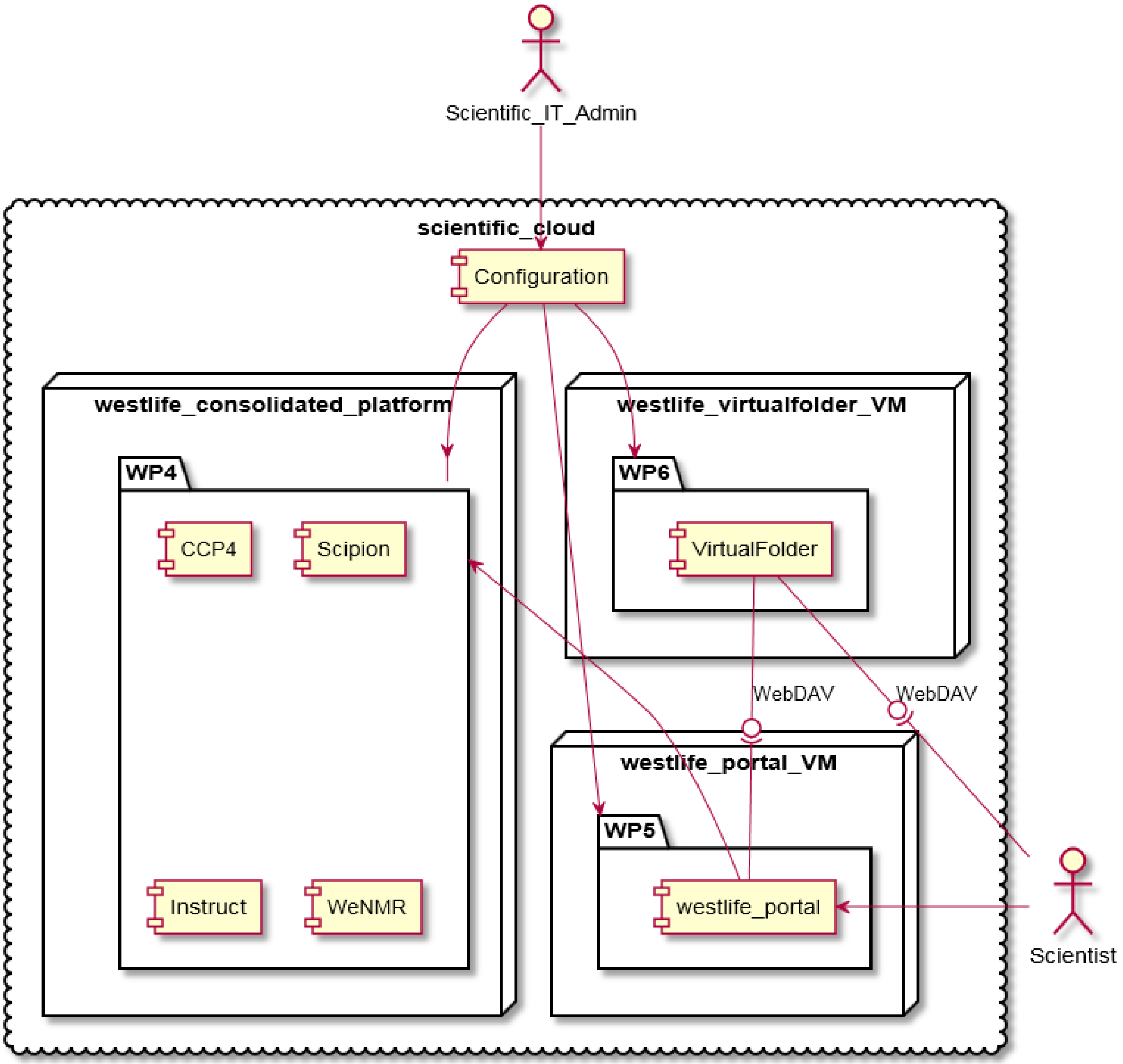


Figure 4: VF deployed as a component of West-Life Portal

In the second deployment scenario, the Virtual Folder works as a component of the West-Life Portal giving appropriate API, protocols and components to access its functionality (see Deliverable 4.1). The diagram (see Figure 4) shows intentional relations among different parts of the system. The scientist interacts mainly with the West-Life portal being developed in WP5. The VF again gives unified access to a range of internal and external data stores, which the user can access directly or via the West-Life portal. In this scenario, scientific software and compute resources for running data analysis and structure solution are separated from the VF.

3.4 Components

A high level view of components in the VF is shown in Figure 5. ControlServices provides a façade pattern [19] to control the mounting/mapping process of individual data storage provided by different providers. WebDAVServer gives a virtual folder view via the WebDAV protocol, so it can be mounted to local resources using network disc mapping feature. WebApp UI is the frontend of the services provided by the VirtualFolder which consist mainly of the FileManager and MiniFileManager components to be reused by the portals of different software.

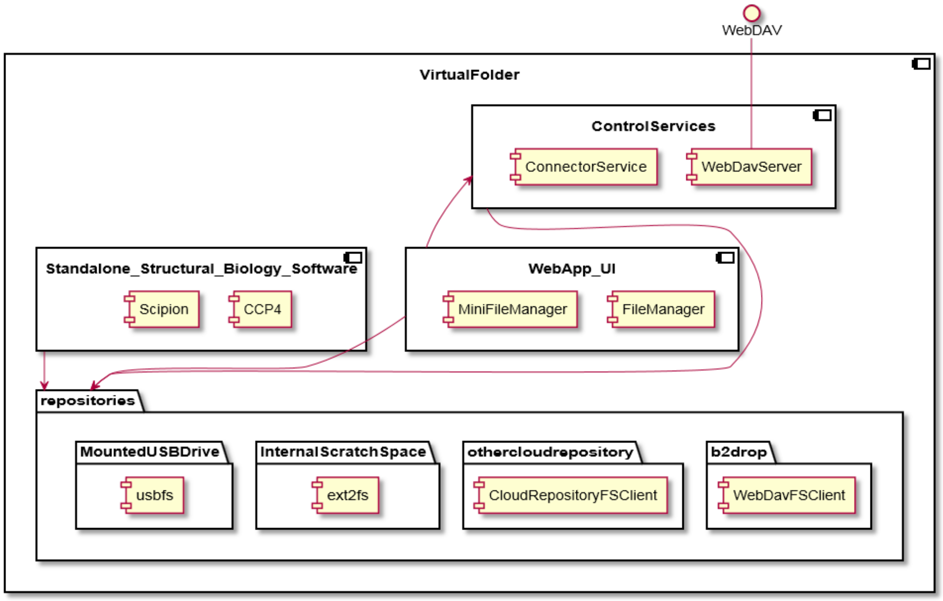


Figure 5: Components of the Virtual Folder

More details are shown in Figure 6. The component HTTP_server gives HTTP-related services including web application in HTML pages, WebDAV module with direct access to repositories and reverse proxy to other related web services. The repositories directory may contain all mountable local or external repositories accessible on the file system level.

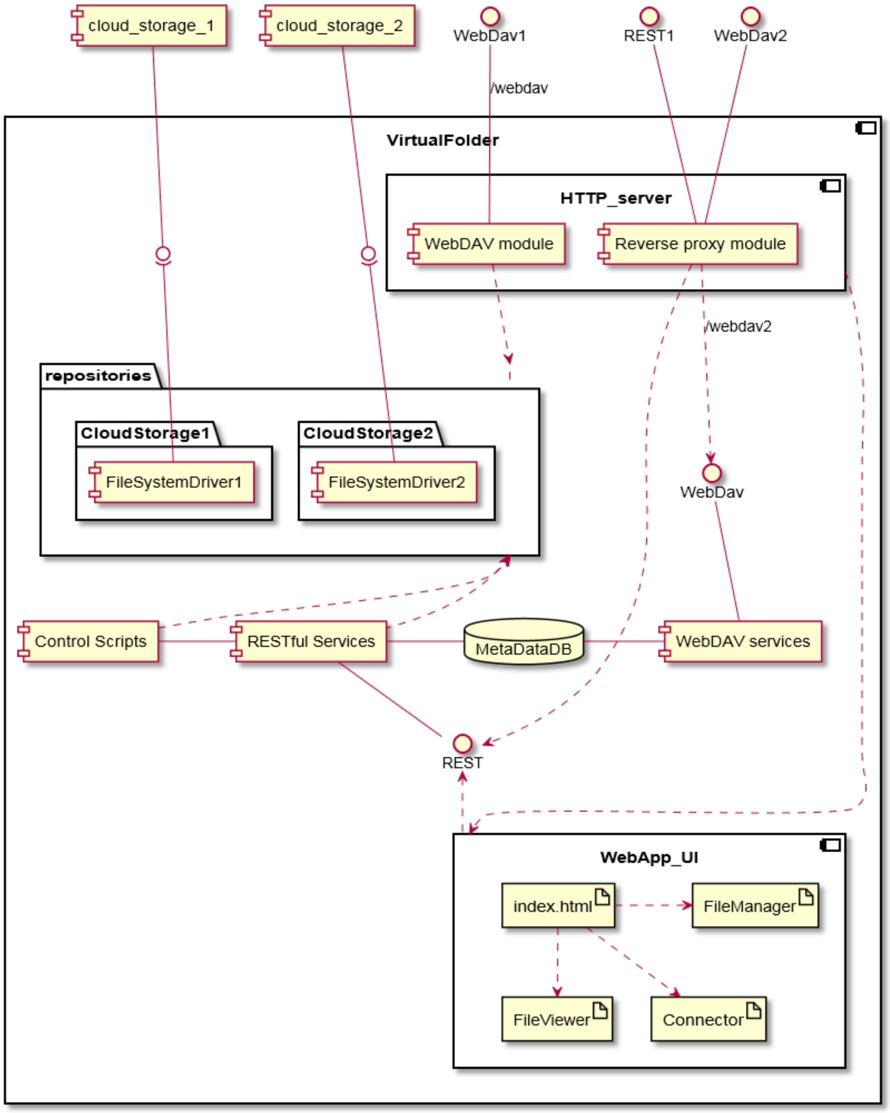


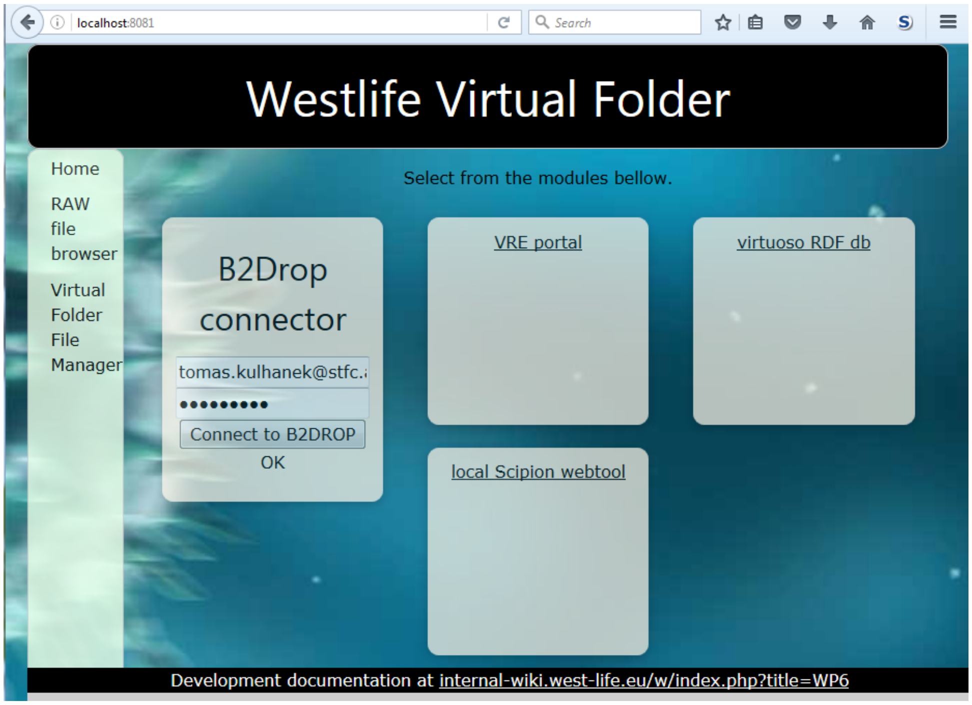
Figure 6: Detailed components of VF

3.5 Prototype Implementation

The single deployment scenario is now implemented, while the integration with the West-life portal will be addressed in further releases. Source codes are published at https://github.com/h2020-westlife-eu/west-life-wp6. Following the instructions in README.md there can be built a single deployment scenario of custom virtual machine with the VF inside it and accessible at http://localhost:8081. Alternatively the VM image is registered at EGI AppDB https://appdb.egi.eu/store/vappliance/d6.1.virtualfoldervm and at https://bio.tools/tool/Virtual_Folder_for_Structural_Biology_Projects/version/16.06.2.

Referring to the example scenario described in Section 3.1,the following is implemented by the virtual folder:

- A scientist can connect the VF with an existing account of a supported repository. To-date, a connection to EUDAT's B2DROP has been implemented.



Westlife Virtual Folder

Select from the modules bellow.

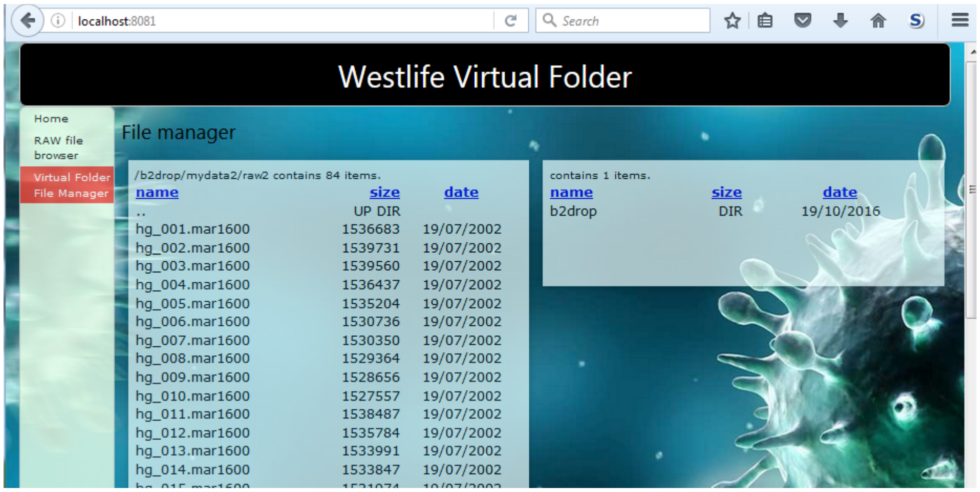
tomas.kulhanek@stfc.a

●●●●●●●●●

OK

Development documentation at internal-wiki.west-life.eu/w/index.php?title=WP6

- The file manager component of VF shows remote files from mounted repository



localhost:8081

Westlife Virtual Folder

Home
RAW file browser
Virtual Folder File Manager

File manager

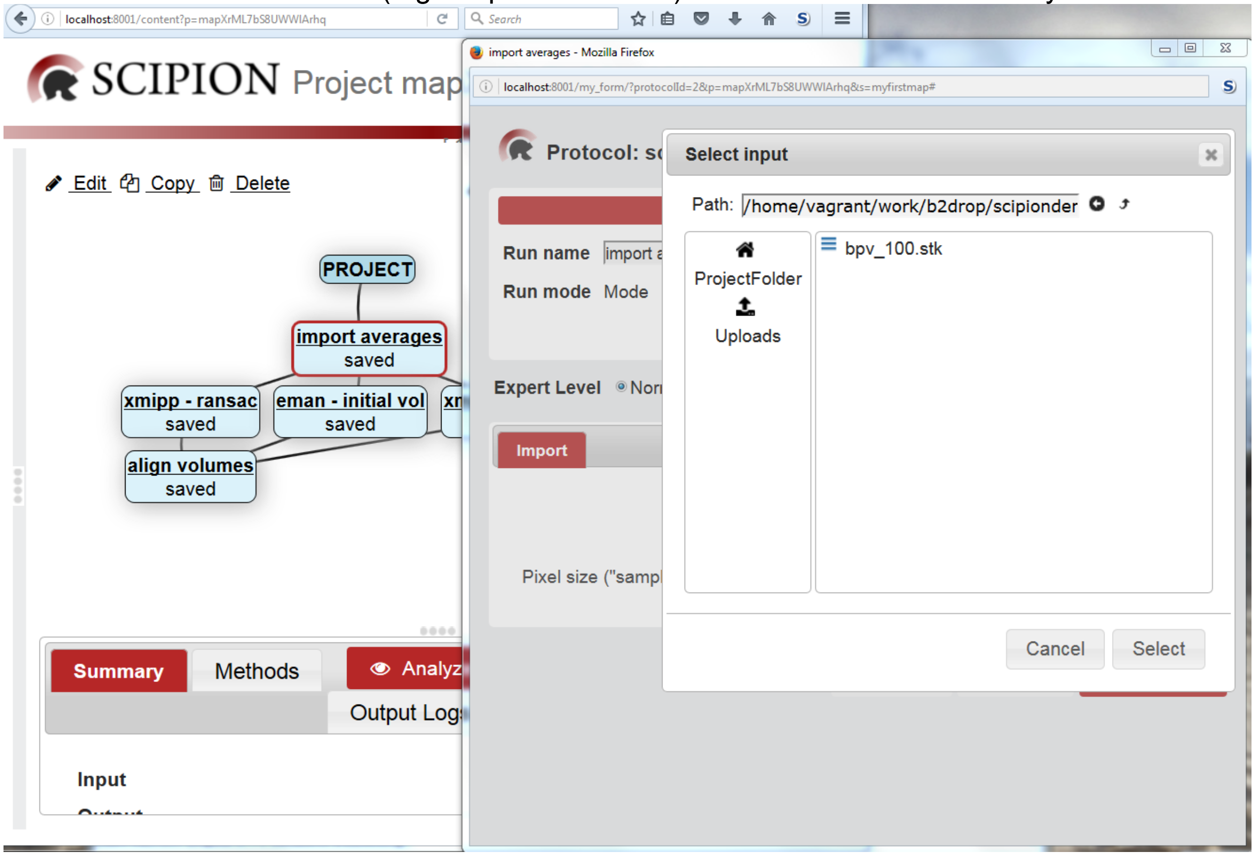
/b2drop/mydata2/raw2 contains 84 items.

name	size	date
..	UP DIR	
hg_001.mar1600	1536683	19/07/2002
hg_002.mar1600	1539731	19/07/2002
hg_003.mar1600	1539560	19/07/2002
hg_004.mar1600	1536437	19/07/2002
hg_005.mar1600	1535204	19/07/2002
hg_006.mar1600	1530736	19/07/2002
hg_007.mar1600	1530350	19/07/2002
hg_008.mar1600	1529364	19/07/2002
hg_009.mar1600	1528656	19/07/2002
hg_010.mar1600	1527557	19/07/2002
hg_011.mar1600	1538487	19/07/2002
hg_012.mar1600	1535784	19/07/2002
hg_013.mar1600	1533991	19/07/2002
hg_014.mar1600	1533847	19/07/2002

contains 1 items.

name	size	date
b2drop	DIR	19/10/2016

- The West-life service (e.g. Scipion Web Tool) can access files mounted by VF



localhost:8001/content?p=mapXrML7bS8UWWIArhq

SCIPION Project map

Edit Copy Delete

Summary Methods Analyz
Output Logs

Input
Output

import averages - Mozilla Firefox

localhost:8001/my_form/?protocolId=2&p=mapXrML7bS8UWWIArhq&s=myfirstmap#

Protocol: sc

Run name import a
Run mode Mode

Expert Level ○ Nor

Import

Pixel size ("sampl

Select input

Path: /home/vagrant/work/b2drop/scipionder

ProjectFolder
Uploads

bpv_100.stk

Cancel Select

In line with the preferred DevOps approach, the Vagrant[4] tool was chosen to provision the development state of virtual machine containing the virtual folder and related components. A generic virtual machine image is taken from public repository hosted at https://atlas.hashicorp.com/westlife-eu. The virtual image contains a minimal installation of the operating system, and contextualization and provisioning is done during the first boot.

Several base OS were tested (Ubuntu, RHEL, Centos, CernVM). The CERN technology was chosen, specifically the micro generic image of CERNVM 4.0 [5], [6] with initial size 17 MB which boots to Scientific Linux 7 (based on RHEL 7) using CernVM-FS[7], [8] technology hosted at cernvm-sl7.cern.ch.

Software distribution is done by CernVM-FS technology too. The repository at west-life.egi.eu was established. It currently hosts west-life specific software and tools including Scipion, Scipion webtools, Virtuoso and a specific build of the MONO implementation of .NET framework. It is mounted to the local file system as a CERNVMFS folder in the virtual machine. Standard proxy and web caching mechanism increase overall performance. This allows rapid startup of the virtual machine and access to the software.

Some other structural software projects are sustained by industrial income, and free only for academic use. Examples are software from CCP4 and CCPN. In these cases the developers are sole distributors of the software. CCP4 has agreed to distribute a virtual machine including their software and the virtual folder. Discussions are ongoing with CCPN. We have communicated to CERN the extra requirements this raises for conditional access to CERNVMFS folders.

The WebDAV protocol is a W3C standard protocol for access to files across the internet. It is offered by many data services, including B2DROP. There are many clients for it, including user friendly ones, allowing the scientist to manage the files collected during the research project. Importantly for our purposes, there are also program-friendly clients for it, including davfs, which allows us to process files accessed this way without any change to the structural biology programs, simply mounting the folder as part of the VM’s file system. However such functionality doesn’t exist for all resource providers (e.g. Dropbox, GoogleDrive, and Amazon S3 give only an API and don’t have official filesystem mounting drivers, although some third-party solution exists).

The Apache open source http server implementation[9] is used with it's mod_proxy serving as reverse proxy to other web related application and mod_dav as WebDAV server points to internal filesystem directory with mounted repositories.

The Virtual Folder imports data in this way and offers an integrated view. It also exports this integrated view as a combined WebDAV folder. The WebDAV services initially uses Milton.io library[17] which implements WebDAV related standard protocol.

The stateless web services provided by the virtual folder follows the RESTful[10] architectural style and utilizes ServiceStack.NET[11] framework. It gives message-based web services based data transfer object and remote façade design patterns[12]. Serialization is done by framework to JSON, XML, JSV based on HTTP header or query format. Database integration is done by ORMLite package [13]. The service implementation can be compiled and executed using MONO implementation of .NET framework on Linux systems[14]. Similar frameworks are available for different platforms, e.g. Jersey for Java[15].

The MetaDataDB database is initialized in PostgreSQL. Any SQL-based DB can be used/replaced in future with no or minimal touch using existing ORM implementation. In the case that a noSQL DB is required, the Virtuoso Universal Database is used as an hybrid database[16].

The WebApp UI is implemented in HTML and following frameworks were tested: Angular JS (version 1), React JS, Aurelia JS.

The webservices provided by the VF implementation are stateless, however, in some situation the stateful web services could be more appropriate, e.g. for following long-term jobs executed on server, cloud etc. This can be addressed by WebSockets technology (HTTP push) and SignalR.NET framework or javax.websocket API.

The VirtualFolder is ready to be integrated into the public West-life portal. Future Web UI direction the following strategies can be addressed.

- Server side rendering – appropriate for complex but single domain web application or portals, pages are more static and should be visible to search engines (e.g. google). Available technologies: Python Django, PHP, ASP.NET, ...

- Client side rendering – appropriate for interactive functionality with dynamical content rendering. Available technologies and frameworks which reduce development and maintenance include: JQuery, Angular, Angular 2, React JS, Aurelia JS, Ember JS, ...

Client side rendering and components seems to be more appropriate for delivering further UI of virtual folder functionality to address point 4) from section 3.0. HTML web component can be neutral to the particular portal backend framework and e.g. PDB component library (http://www.ebi.ac.uk/pdbe/about/news/introducing-pdb-component-library) is following similar strategy.

The WebApp UI components are designed using standard HTML (version 5), Javascript frameworks and communicating with server using AJAX techniques[3]. The FileManager web component frontend and it's relation to RESTful web service are schematized in Figure 7.

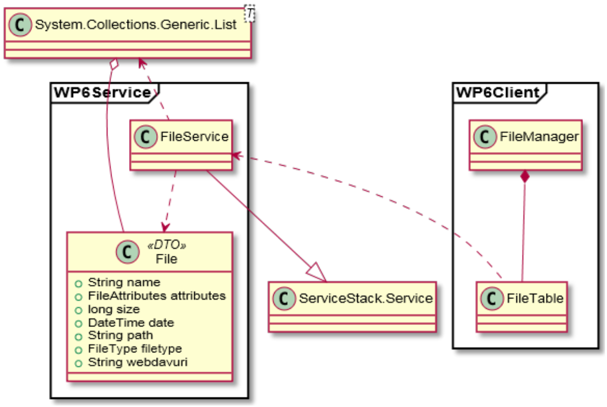


Figure 7. Class diagram of decoupled Service backend and UI components frontend

The current implementation keeps the user's credential of B2DROP account within the VM as there is no other authentication mechanism offered by EUDAT. Such user credentials can be exploited. An access token mechanism (e.g. OAuth, OpenID) would be more sufficient and has been already delivered e.g. by the EUDAT's B2SHARE service.

References cited

BioRegistery http://bioregistry.cbs.dtu.dk/index.html
CERIF http://www.eurocris.org/ontologies/cerif/1.3
EDAM http://edamontology.org/page
PROV-O http://www.w3.org/TR/prov-o/
VoID http://www.w3.org/TR/void/
Well-Known URIs https://tools.ietf.org/html/rfc5785

[1] “File Manager.” [Online]. Available: https://en.wikipedia.org/wiki/File_manager. [Accessed: 27-Sep-2016].

[2] “Orthodox file managers standard.” [Online]. Available: http://www.softpanorama.org/OFM/Standards/index.shtml. [Accessed: 27-Sep-2016].

[3] “Asynchronous Javascript and XML (AJAX).” [Online]. Available: https://en.wikipedia.org/wiki/Ajax_(programming). [Accessed: 27-Sep-2016].

[4] “Vagrant.” [Online]. Available: https://www.vagrantup.com/. [Accessed: 21-Sep-2016].

[5] “CernVM.” [Online]. Available: http://cernvm.cern.ch/. [Accessed: 21-Sep-2016].

[6] J. Blomer, D. Berzano, P. Buncic, I. Charalampidis, G. Ganis, G. Lestaris, R. Meusel, and V. Nicolaou, “Micro-CernVM: Slashing the Cost of Building and Deploying Virtual Machines,” *J. Phys. Conf. Ser.*, vol. 513, 2014.

[7] “CernVM-FS.” [Online]. Available: https://cernvm.cern.ch/portal/filesystem. [Accessed: 21-Sep-2016].

[8] J. Blomer, P. Buncic, R. Meusel, G. Ganis, I. Sfiligoi, and D. Thain, “The Evolution of Global Scale Filesystems for Scientific Software Distribution,” *Comput. Sci. Eng.*, vol. 17, no. 6, pp. 61–71, 2015.

[9] “Apache HTTP server.” [Online]. Available: https://httpd.apache.org/. [Accessed: 21-Sep-2016].

[10] R. T. Fielding, “Chapter 5: Representational State Transfer (REST),” *Archit. Styles Des. Network-based Softw. Archit. Diss.*, 2000.

[11] “ServiceStack.” [Online]. Available: http://servicestack.net/. [Accessed: 21-Sep-2016].

[12] M. Fowler, *Patterns of Enterprise Application Architecture*. Pearson Education, Inc., 2003.

[13] “ServiceStack OrmLite.” [Online]. Available: https://github.com/ServiceStack/ServiceStack.OrmLite. [Accessed: 21-Sep-2016].

[14] “MONO project.” [Online]. Available: http://www.mono-project.com/. [Accessed: 21-Sep-2016].

[15] “Jersey - RESTful Web Services in Java.” [Online]. Available: https://jersey.java.net/. [Accessed: 21-Sep-2016].

[16] “Virtuoso Universal Database.” [Online]. Available: http://virtuoso.openlinksw.com/. [Accessed: 21-Sep-2016].

[17] “Milton.io Java WebDAV library.” [Online]. Available: http://milton.io/. [Accessed: 21-Sep-2016].

[18] "Using OAuth 2.0 to Access Google APIs." [Online]. https://developers.google.com/identity/protocols/OAuth2

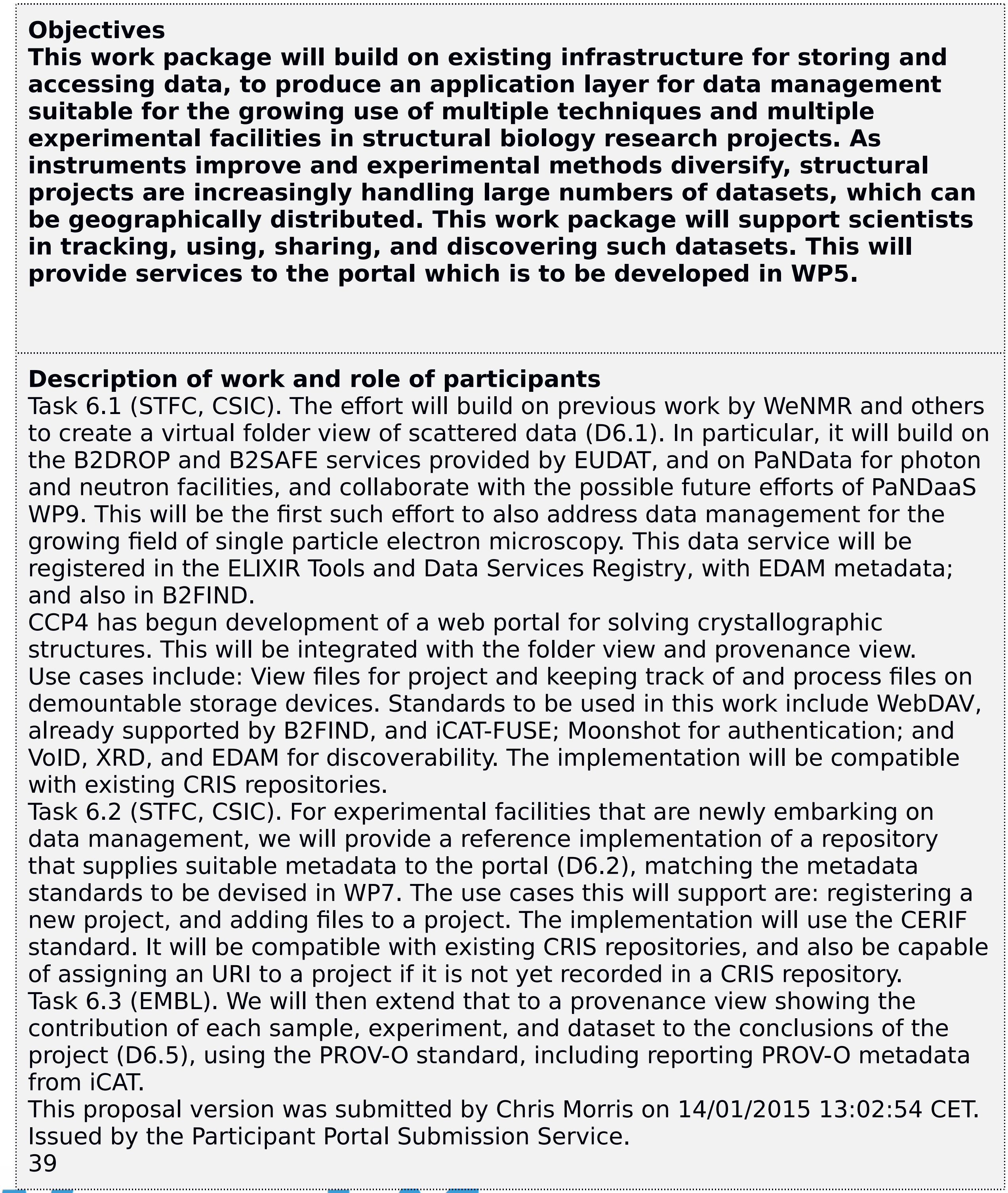
[19] Gamma, Erich, et al. "Design patterns: Elements of reusable software architecture." Addison-Wesley (1995).

Delivery and Schedule

Unfortunately there were some delays in hiring a developer to work on this deliverable. Tomas Kulhanek began work in March 2017. Most of the planned deliverable is complete, as reported on above. Future work will integrate users' demountable storage devices into the virtual file system.

The original plan for this deliverable also included collaboration with PaNDaaS, if that grant were to be awarded. In the event, it was not. The iCAT data service does not offer a full WebDAV interface, Nevertheless, it is very desirable that datasets in iCAT at Diamond Light Source and other facilities are available in the West-Life virtual folder system. This work will be performed, but could not be completed in time to become part of this deliverable.

Background information



Objectives

This work package will build on existing infrastructure for storing and accessing data, to produce an application layer for data management suitable for the growing use of multiple techniques and multiple experimental facilities in structural biology research projects. As instruments improve and experimental methods diversify, structural projects are increasingly handling large numbers of datasets, which can be geographically distributed. This work package will support scientists in tracking, using, sharing, and discovering such datasets. This will provide services to the portal which is to be developed in WP5.

Description of work and role of participants

Task 6.1 (STFC, CSIC). The effort will build on previous work by WeNMR and others to create a virtual folder view of scattered data (D6.1). In particular, it will build on the B2DROP and B2SAFE services provided by EUDAT, and on PaNData for photon and neutron facilities, and collaborate with the possible future efforts of PaNDaaS WP9. This will be the first such effort to also address data management for the growing field of single particle electron microscopy. This data service will be registered in the ELIXIR Tools and Data Services Registry, with EDAM metadata; and also in B2FIND.

CCP4 has begun development of a web portal for solving crystallographic structures. This will be integrated with the folder view and provenance view.

Use cases include: View files for project and keeping track of and process files on demountable storage devices. Standards to be used in this work include WebDAV, already supported by B2FIND, and iCAT-FUSE; Moonshot for authentication; and VoID, XRD, and EDAM for discoverability. The implementation will be compatible with existing CRIS repositories.

Task 6.2 (STFC, CSIC). For experimental facilities that are newly embarking on data management, we will provide a reference implementation of a repository that supplies suitable metadata to the portal (D6.2), matching the metadata standards to be devised in WP7. The use cases this will support are: registering a new project, and adding files to a project. The implementation will use the CERIF standard. It will be compatible with existing CRIS repositories, and also be capable of assigning an URI to a project if it is not yet recorded in a CRIS repository.

Task 6.3 (EMBL). We will then extend that to a provenance view showing the contribution of each sample, experiment, and dataset to the conclusions of the project (D6.5), using the PROV-O standard, including reporting PROV-O metadata from iCAT.

This proposal version was submitted by Chris Morris on 14/01/2015 13:02:54 CET. Issued by the Participant Portal Submission Service.

39

West-Life
39
Task 6.4 (EMBL): We will facilitate collaboration between the computational modeling and structural biology community by providing the necessary data management, dissemination and analysis tools (D6.3) for predicted models. We use metadata standards to be developed in WP7 for description of predicted models, and implement the necessary data dissemination mechanisms (D6.4).

D6.1 Report on the implementation of the folder view: a virtual file system view of scattered data - STFC

D6.2 Report on the reference implementation of a repository providing CRIS and PROV-O metadata suitable to support the folder view, provenance view, and data processing facilities. - STFC

D6.3 Report on the search and query interface for modeled assemblies (including CAPRI models) – EMBL-EBI/PDBe

D6.4 Report on the implementation of the provenance view: a dependency graph of samples and dataset. – STFC.

References

BioRegistery http://bioregistry.cbs.dtu.dk/index.html
CERIF http://www.eurocris.org/ontologies/cerif/1.3
EDAM http://edamontology.org/page
PROV-O http://www.w3.org/TR/prov-o/
VoID http://www.w3.org/TR/void/
Well-Known URIs https://tools.ietf.org/html/rfc5785

Deliverables

	Title	Lead	Type	Dissemination	Due
D6.1	Report on Virtual folder	STFC	R	PU	M12
D6.2	Repositories	STFC	DEM	PU	M24
D6.3	Report on Assembly queries	EMBL	R	PU	M32
D6.4	Report on Provenance	STFC	R	PU	M33