Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	UniProt
Website:	https://www.uniprot.org/
Certification Date:	15 May 2020

This repository is owned by: **EMBL-EBI**

UniProt

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Other (Please describe below)

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of Repository

Knowledgebase

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The UniProt consortium and host institutions EMBL-EBI, SIB and PIR are committed to the long-term preservation of the UniProt databases.

UniProt is a knowledgebase in the field of molecular biology that contains protein sequences and annotations. We use the term knowledgebase as we accumulate, organize, and link growing bodies of information related to core datasets using significant curation efforts beyond those of a domain repository. This is the same knowledgebase definition as provided by the NIH Strategic Plan for Data Science (https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf)

The large majority of the protein sequences are derived from translation of the holdings of the worldwide consortium of DNA sequence databases (INSDC repositories; http://www.insdc.org). Our team adds value to the protein sequences through expert curation and automated annotation systems to give both depth and breadth of annotation. UniProt is the domain repository for submission of direct protein sequencing experiments as recommended, among others, by the Nature Publishing Groups Journal Scientific Data (https://www.nature.com/sdata/policies/repositories). We accept direct submissions of protein sequences determined at the protein level only via SPIN (https://www.ebi.ac.uk/swissprot/Submissions/spin/account/login). However, these user-submitted sequences account for less than 0.1% of the total sequences in the resource. We regularly publish peer reviewed articles in the scientific literature to describe new features and updates. In particular, the update articles we publish in the Nucleic Acids Research Database Issue every year or two are extremely highly cited.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

The UniProt resource is used by a diverse set of user communities. Overall, we are accessed via the web by 700,000 unique users per month according to Google Analytics. These users are primarily life science and biomedical researchers from both academia and industry. One can get a feeling for the diversity of users' subject areas from the graph below, which shows the set of EMBL-EBI users that self-report as users of UniProt (data derived from the 2015 EMBL-EBI user survey).

FIGURE 1 (Primary field of study of self-reported UniProt users) here

The UniProt resource is composed of three core databases: The primary resource is the UniProt Knowledgebase (UniProtKB), additional related and derived datasets are the UniProt Reference Clusters (UniRef; https://www.uniprot.org/help/uniref) and the UniProt Archive (UniParc). The UniProt Knowledgebase is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. It consists of two sections: UniProtKB/Swiss-Prot is an expertly curated, non-redundant protein database that aims to provide all known relevant information about experimentally characterized proteins. The UniProtKB/TrEMBL section contains a comprehensive and high-quality collection of protein sequences, which are enhanced through computational analyses. UniParc is a non-redundant database which contains all available protein sequences from the publicly available sequence databases. The UniRef clusters (UniRef100, UniRef90 and UniRef50) consist of sets of protein sequences from UniProtKB and selected UniParc records clustered at 3 levels of sequence identity (50,90 and 100%). See graphic and help documentation on website (https://www.uniprot.org/help/about). Also see a description of our history (https://www.uniprot.org/help/about) and current funders.

The UniProt databases are central to the activities of many other resources around the world. UniProt acts as a provider of annotation, nomenclature, cross-references to other resources and the protein sequences themselves. UniProt provides cross-links to 150 molecular biology resources and for some of these, the traffic from UniProt cross-references provides a large fraction of their traffic.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation, D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

The level of curation in UniProtKB consists of levels B-D for both the Swiss-Prot and TrEMBL sections of our main database. All imported data undergoes initial quality checks, some format conversion and may undergo some automated annotation. Curation goes well beyond D in many cases. In the UniProtKB/Swiss-Prot section all entries are reviewed by expert curators and there is ongoing literature-based curation. In the TrEMBL section various automated rules, both human curated and machine generated, add additional functional annotation.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

We do not outsource any of our core activities. Consortium members may outsource some IT related activities such as virtual servers but otherwise manage our infrastructure. Currently PIR and SIB use internal institutional resources but have investigated some cloud service providers such as AWS and Google as future options.

EMBL-EBI has contracted with Kao Data services in the UK to deliver data center services

https://www.computerweekly.com/news/252470579/Kao-Data-secures-EMBL-EBI-as-tenant-as-bid-to-become-leading-life-sciences-colo-gathers-pace

Reviewer Entry

Reviewer 1

Comments:

Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

Collaborations: We collaborate and partner with related resources and import sequence data and initial metadata from other well established high quality sequence repositories and genome annotation resources including: EMBL (https://www.ebi.ac.uk/ena/about)/GenBank/DDBJ, Ensembl (http://www.ensembl.org/info/about/index.html), RefSeq (https://www.ncbi.nlm.nih.gov/refseq/about/), PDB (http://www.wwpdb.org), Model Organism Databases (https://www.alliancegenome.org), OMIM (https://www.omim.org) and others.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The UniProt mission statement is:

“The mission of UniProt (https://www.uniprot.org) is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information”. The statement above is at the top of our home page and this explicit mission is the one for which our funders support us. This mission and related goals is explicitly stated in all our applications for funding and the main grant that funds us is entitled “UniProt: A centralized protein sequence and function resource”. (https://taggs.hhs.gov/Detail/AwardDetail?arg_AwardNum=U41HG007822&arg_ProgOfficeCode=55). The UniProt resource and its forerunners have been operating for over 30 years. We have public copies of previous releases going back to UniProt 1.0 in 2003. Earlier releases of the founding databases SWISS-PROT, TrEMBL and PIR Protein Sequence Database are available from the founding institutions (described below in R5).

The UniProt consortium and our host institutions at EMBL-EBI, SIB and PIR are committed to the long-term preservation of the UniProt data in the unlikely case that funding for our current activities for this core resource is lost without a successor institution. Please see this statement on the top of our website page about UniProt https://www.uniprot.org/help/about

“The UniProt consortium and host institutions EMBL-EBI, SIB and PIR are committed to the long-term preservation of the UniProt databases.”

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All UniProt data is freely available under the Creative Commons Attribution 4.0 International license. Please see https://www.uniprot.org/help/license and https://creativecommons.org/licenses/by/4.0/. This license allows users to 1) copy and redistribute the material in any medium or format, 2) adapt remix, transform, and build upon the material for any purpose, even commercially, as long as 3) the user gives appropriate credit, provides a link to the license, and indicates if changes were made. Users may do so in any reasonable manner, but not in any way that suggests the licensor endorses them or their use.

Our license comes with the following disclaimer:

We make no warranties regarding the correctness of the data, and disclaim liability for damages resulting from its use. We cannot provide unrestricted permission regarding the use of the data, as some data may be covered by patents or other rights.

Any medical or genetic information is provided for research, educational and informational purposes only. It is not in any way intended to be used as a substitute for professional medical advice, diagnosis, treatment or care.

We have reviewed the licenses and terms of use for our source databases where they exist and nothing prevents us from our redistribution with appropriate credit. Links to the appropriate usage policies of these databases are provided below.

INSDIC databases: https://www.ncbi.nlm.nih.gov/home/about/policies/#data , https://www.ebi.ac.uk/about/terms-of-use , https://www.ddbj.nig.ac.jp/insdc-e.html#policy
PDB: ftp://ftp.wwpdb.org/pub/pdb/advisory.txt
OMIM: https://www.omim.org/help/agreement

However, we also use software for predictions of sequence features, which are annotated and distributed. This software is free to academics but some are licensed for commercial use. Our data is free to commercial users so there is some possibility that this is a violation we are not aware of, so we are beginning a process to check with the original software developers about this.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

UniProt has an 18-year track record for providing continuous support for open access to data and services. Since 2002 it has been managed by a global partnership that collaborates on database production and annotation which further strengthens the long-term stability of the organisation. Prior to 2002 the individual consortium partners European Molecular Biology Laboratory- European Bioinformatics Institute (EMBL-EBI; https://www.ebi.ac.uk), the SIB Swiss Institute of Bioinformatics (SIB; https://www.sib.swiss) and the Protein Information Resource (PIR; https://proteininformationresource.org) have managed and distributed their own protein databases for 20 years or more.

UniProt is principally funded by three major sources: National Institutes of Health (NIH), The European Molecular Biology Laboratory (EMBL) and The Swiss Secretariat for Education, Research and Innovation (SERI) (https://www.uniprot.org/help/about). The grant funding horizons for these partners vary between 3 to 5 years and have staggered renewal dates. Thus, our funding portfolio is diversified across funders as well as geographically. This greatly reduces risks of total funding loss.

The UniProt resource is one of the most important knowledgebases for molecular biology. UniProt was one of the 19 resources selected as an ELIXIR Core Data Resource (https://www.elixir-europe.org/platforms/data/core-data-resources)

in 2016, based on a set of agreed selection criteria (PMID: 27803796). UniProt has also taken part in meetings of the Global Biodata Coalition, a coalition of funders in the life sciences that aim to sustain the funding of data resources. Recognition from both ELIXIR and the Global Biodata Coalition as well as the Core Trustworthy Data Repository help to highlight the importance of protein sequence and function annotation provision to funders.

One risk associated with data resources is the retirement of key staff. The UniProt management structure has demonstrated great resilience over the decades with multiple PI changes over the last two decades. Each Institution has managed transitions of PIs to ensure business continuity. These transitions have been discussed with the other partners to ensure a strong working relationship is maintained. At the Swiss Institute of Bioinformatics, the leadership has changed from the founder Amos Bairoch to Ioannis Xenarios and most recently to Alan Bridge. At EMBL-EBI, leadership has changed from Rolf Apweiler to Alex Bateman.

The host Institutions have very long histories of data provision in the life sciences. Each freely provides technical and administrative infrastructure that greatly reduces the funding burden for the resource. The host Institutions would be able to bridge short losses or reductions in funding to enable retention of staff. Key staff across the sites are largely Institutionally funded, ensuring that critical management functions are protected in the case of funding losses.

In the case of loss of funding to one of the UniProt partners followed by continued lack of funding, their critical activities would be passed on to the remaining partners to ensure the smooth running of the UniProt knowledgebase. This transitioning would be repeated with funding losses by other partners. In the extremely unlikely case of total loss of funding to all partners there are several mitigation strategies:

- Data is currently released under a CC-BY license which allows the UniProt data to be used by anyone to create a new derivative protein resource.
- Final data distributions will be maintained by the host Institutions in perpetuity.
- Sequence data is available via other resources such as RefSeq, INSDC resources, Ensembl etc.
- We would hand over data and software to a funded third party after evaluation in the event of sunsetting the UniProt project

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The data collected and distributed by UniProt are considered public data and do not present ethical disclosure risks. No personally identifiable information is collected or maintained except for authors names on publicly available manuscripts or sequence submissions. For website usage we do not collect or track user identifiable information and conform the legal requirements of the European Union, United States of America and Switzerland. Complying with the most stringent requirements in the cases where they differ. UniProt staff are bound by the ethics guidelines of their home institution on professional conduct.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

UniProt is an essential component of the biomedical research infrastructure and is supported by a large dedicated team. The ongoing success of UniProt relies on effective leadership and management structures underpinned by our core values of scientific integrity, quality, comprehensiveness, stability, openness and collaboration. Effective management of UniProt will ensure communication, productivity, data quality and data integration as well as the continuous evaluation of new data sources and modern technologies, thus improving efficiency and value of the resource for the community. The UniProt resource is run by the UniProt Consortium, founded in 2002, which consists of three partners with comprehensive experience in the handling of protein sequence databases who have worked together successfully for the last 17 years.

Organizational structure and staff responsibilities: Each group of the consortium is managed individually and the Principal Investigator (PI), Dr Alex Bateman at the EMBL-European Bioinformatics Institute (EMBL-EBI) in Cambridge, UK, handles the overall project management and coordination. At the Protein Information Resource (PIR) at both Georgetown University and the University of Delaware at Newark, Prof Cathy Wu acts as UniProt PI, and at Swiss-Prot at the SIB Swiss Institute of Bioinformatics (SIB) in Geneva, Switzerland, the PI is Dr Alan Bridge. The PIs are in close regular contact to discuss project progress, overall project directions and priorities, resource allocation and budget issues. A team of senior staff members assists the PIs with the day to day management, and the project is overseen by the Scientific Advisory Board (SAB) that has been appointed by the UniProt Consortium to provide independent scientific advice.

Generally, interaction between consortium partners is primarily carried out using electronic communication channels. Project quality and the monitoring of progress are ensured through the following measures:
● The Scientific Advisory Board (SAB) meets annually to evaluate the progress of the database and its overall responsiveness to user needs.
● The UniProt Consortium, through the PI, works closely with the NHGRI program director to keep priorities under review and to evaluate progress in meeting the research community's needs.
● UniProt Consortium meetings provide an opportunity to present developments, follow up on milestones, and to discuss, in detail, any unresolved issues. These three-day meetings take place on a bi-annual basis rotating between the sites. The consortium favors collective decision-making, but, if necessary, there will be voting on unresolved questions with one vote for each partner.

● Regular tri-weekly strategy conference calls take place among the key personnel of all partners, chaired by the PI, to exchange ideas, report progress and identify problems.
● Regular monthly operations conference calls to discuss operational aspects of software development and annotation to ensure standards are developed and maintained across the different sites.
● A project wiki, issue tracking system, and e-mail lists are used to exchange progress reports, discuss issues, document procedures and troubleshoot. This allows open and fast exchange of information and fast decision-making.

Visits between the scientific and technical staff from the different sites take place whenever appropriate. These allow staff members to get to know each other, to get familiar with each sites' work projects, and to solve scientific and technical problems.

Funding: Currently, funding for the UniProt project comes from the NIH and several European sources. The European Molecular Biology Laboratory (EMBL) and the Swiss State Secretariat for Education, Research and Innovation (SERI) are the main funders of the UniProt activities at EMBL-EBI and SIB. The NIH is the main funder of the UniProt activities at PIR and contributes approximately 40% of the full cost of UniProt activities at EMBL-EBI and SIB. Overall the project has ~80 FTE staff of which approximately half are biocurators see table 1 below for breakdown of current FTEs. FTE's do vary some over time. Current funding sources are acknowledged here (https://www.uniprot.org/help/about). Experienced "Key Staff" help manage and coordinate activities at each staff. Names and descriptions of Key Staff are found here: (https://www.uniprot.org/help/key_staff).

Category Total
Biocurators [FTEs] 42.5
Developers [FTEs] 33
User support and communication [FTEs] 2
Management [FTEs] 5
Total FTEs 82.5

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific

guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The UniProt Consortium has appointed a Scientific Advisory Board (SAB) to provide independent scientific advice regarding the project. The UniProt SAB provides guidance to the PIs on all aspects of the project, including scientific and technical elements impacting on the specification and maintenance of UniProt and on the roles of the UniProt partners. The SAB aids the UniProt Consortium in making decisions by reviewing progress, offering independent expert judgment, and highlighting issues likely to be of future concern that lie within their terms of reference. Specifically, the role of the SAB members is to:

- Represent the broad interests of the life science community.
- Advise on all scientific aspects of the project.
- Assess the effectiveness and suitability of inter-consortium communication practices.
- Assess progress towards meeting milestones and deliverables, and to advise on changes to prioritization of these objectives.
- Advise on short, medium and long-term scientific opportunities, risks, as well as cost-effectiveness.

The SAB currently consists of 12 members with one member appointed as chair. The current and past SAB membership can be found on the UniProt website here (https://www.uniprot.org/help/sab). This number allows for a satisfying turnout at SAB meetings. The advisors are carefully selected to represent a broad spectrum of academic and industrial interests and comprise distinguished experts as well as slightly less senior scientists. Members rotate on a four-year basis, which allows for changes in the SAB's composition according to changing trends in the field. For example, our recent recruitments have been in the areas of clinical informatics, drug discovery and pharmacogenomics to represent our increasing clinical relevance.

Meetings of the SAB with key UniProt personnel take place once a year with circulation of an agenda and supporting documentation at least two weeks in advance. The agenda is prepared in consultation with the SAB chairperson following

consultation with the key UniProt staff to identify important areas requiring independent review and advice. Following each meeting, the SAB chairperson coordinates the production of a report outlining the advice of the committee which will be acted on by the UniProt Consortium. This mechanism has proved to be a great asset. Additional ad-hoc teleconferences may be organized if necessary, to address specific issues that cannot wait for the next scheduled meeting. The SAB has been instrumental in many important decisions such as the decision to wind down UniMES the metagenomic portion of UniProt, as well as shaping our strategy for selecting reference proteomes using the community. A detailed study of the sustainability of literature curation (PMID: 29036270) was from a direct suggestion of our SAB. Our goal is to continue working closely with our scientific advisors to benefit from external independent advice and to ensure continued improvements.

In addition to the UniProt SAB, the project also undergoes independent reviews at both EMBL-EBI on a four yearly cycle and at the SIB on a five-yearly cycle. These Institutional reviews are carried out by panels of eminent scientists and are used to determine the future support from EMBL and SIB.

As well as our SAB we have many diverse interactions with our user community. These range from interactions through our helpdesk, our in-person training courses, webinars, conference presentations and social media channels. Through to participation in specific workshops that we host and taking part in targeted curation meetings. One particularly fruitful method of communication is through User Experience testing of our web developments. Here we have detailed interviews with specific target users to understand how best to develop features on the website.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

UniProt maintains a regular release cycles for all its databases, currently making 10 releases each calendar year. Each release has release notes on new information added any format or annotation changes made since a previous release. For example, see www.uniprot.org/news/2019/01/16/release . Users are also notified of forthcoming format changes. See www.uniprot.org/help/?fil=section:changes . Statistics on each release are also provided www.uniprot.org/statistics/?sort=published .

UniProt also distributes data via its FTP site ftp.uniprot.org and maintains all previous releases back through 2005 ftp.uniprot.org/pub/databases/uniprot/previous_releases/. Release metalink files contain specific information file information including location, size and checksums for each data file in a directory. See ftp.uniprot.org/pub/databases/uniprot/current_release/RELEASE.metalink for an example.

Upon integration into UniProtKB, each entry is assigned a unique accession number, called the ‘Primary (citable) accession number’. Entries are accessed by accession numbers via a standard URL (www.uniprot.org/uniprot/P68250). Entries can have more than one accession number. This can be due to two distinct mechanisms: a) When two or more entries are merged, the accession numbers from all entries are kept. The first accession number is referred to as the ‘Primary (citable) accession number’, while the others are referred to as ‘Secondary accession numbers’; b) If an existing entry is split into two or more entries (‘demerged’), new ‘primary’ accession numbers are attributed to all the split entries while all original accession numbers are retained as ‘secondary’ accession numbers. Occasionally obsolete sequences are removed from the primary databases but the sequences and their accessions remain in the UniParc Archive (https://www.uniprot.org/help/uniparc) which contains protein sequences past and present from all sources. Details and examples of primary and secondary accessions and the difference between accessions and entry names can be found on our help pages here (www.uniprot.org/help/accession_numbers).

The history of annotation changes is tracked in UniSave (PMID: 16551660; http://www.ebi.ac.uk/uniprot/unisave/app/#/) and available for each entry on the website. For example, www.uniprot.org/uniprot/O14733?version=*

UniProt distributes its data in multiple data formats and schemas are available at: www.uniprot.org/docs/?query=schema and sparql.uniprot.org .

Additional documentation on data management and other related topics is available on the UniProt help pages (https://www.uniprot.org/help/).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Import of directly sequenced proteins.
UniProt accepts submissions of directly sequenced proteins where the sequence has been obtained by Edman degradation or by manual interpretation of MS/MS spectra. An online submission tool, SPIN (http://www.ebi.ac.uk/swissprot/Submissions/spin/), is provided which ensures collection of all necessary data. A curator checks all submitted data including the methodology used to generate the sequence data and rejects submissions which do not meet the criteria for acceptance (https://www.ebi.ac.uk/swissprot/Submissions/spin/help). If data are suitable for acceptance, a curator corresponds with the submitter to clarify any inconsistencies or other issues with the submitted data. Once the curator is happy with the submitted data, a UniProt record is created and the unique accession number for the record is provided to the submitter for use in publications. This service enables researchers to make their de novo–sequenced proteins available to the scientific community and acquire UniProt accession numbers for use in publications. This submission route provides less than 0.1% of our current protein sequence holdings.

Import of sequences derived from DNA translations.

The vast majority protein sequences come from translations of DNA coding regions defined by the submitter of the DNA sequence. We import sequences and metadata directly from appropriate nucleotide and genome repositories including EMBL-NA, Ensembl, RefSeq, PDB and others. We do this with in collaboration with the source repositories with specific filters and QC checks in place. The outcome of the import process is that all new protein sequences are added to the UniProt archive which contains all protein sequences from all sources. A subset of UniParc sequences go into UniProtKB/TrEMBL if they meet our criteria for being naturally occurring DNA encoded protein sequences from a known source, are not from patent databases or contain sequences artifacts derived from recombinant laboratory cloning and are not sequences previously determined to be inappropriate for inclusion by UniProt or the scientific community (i.e. contaminated sequences, mis-identified sequences or organisms, translations from community recognized low quality gene calling methods). New data is imported into TrEMBL unless it already exists in Swiss-Prot. Data imported into Swiss-Prot are manually selected and reviewed by an expert curator.

All new sequences are assigned a unique permanent accession number and links to the original sources accession numbers and databases are maintained. Some annotations from the original source are maintained and flagged as imported from the source. If at some point a UniProtKB sequence is removed from TrEMBL or Swiss-Prot for some reason its sequence and accession remain in UniParc along with links to its original source.

Data is provided to users in multiple formats with documentation. Some of these formats are established community defined formats (i.e. FASTA, GFF, XML, RDF) some like our legacy text-based flat file format have become standards that are supported by other systems. All formats are documented on the UniProt website.

MetaData Requirements.
For direct protein sequence submission, the metadata requirements are defined here: https://www.ebi.ac.uk/swissprot/Submissions/spin/help. Protein sequences from translations of DNA coding regions come with metadata defined by the INSDC databases like EMBL-ENA which has standards and a metadata model (https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html). UniProt contains well over 100 data types (see https://www.uniprot.org/owl/core.rdf) and there are only few that can be covered with standard metadata schemes like the Dublin Core (for bibliography), or community vocabularies like the Sequence Ontology and FALDO (for the description of sequence ranges). UniProt is actively engaged in the developments of these domain-specific vocabularies to extend them and increase their adoption.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All three sites maintain separate backups of the public releases of the UniProt datasets including the FTP site and Website. The backups include all public versions for UniProtKB from its first release in 2003. In addition, the sites backup their internal databases, software, related files used in their production pipelines and process documentation.

EMBL-EBI's state-of-the-art technical architecture is secure and robust, and data is distributed in three discrete data centres in different geographical locations to assure long-term security. This gives our data very high protection through redundancy and provides sufficient capacity and reserve to ensure our management of the rising influx of data and compute requests. Our storage solution can scale vertically and horizontally (scale-out). Resources to support the storage solution can be added to each individual node, and nodes can be added to the entire solution.

PIR and SIB backups include: Redundant servers, Mirroring critical disk storage systems; Tape backups stored off site; and archiving older data using Apache Hadoop.

Data is transferred between sites using Aspera (https://asperasoft.com) and MD5 checksums are compared. For users downloading from FTP sites we provide metalink files with MD5 checksums for all downloadable files, for example ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/RELEASE.metalink

Processes and SOPs are documented using a Confluence workspace (https://www.atlassian.com/software/confluence) for sharing documentation. Development Issues are tracked using Jira (https://www.atlassian.com/software/jira). These services are included in all backups.

In the event that a web server goes down users are automatically redirected to another mirror site while a backup server is

brought online. If an FTP site goes down users will be redirected to a backup server or one of the other consortium institutions FTP servers. System admins at each site have access to the UniProt.org DNS if emergency changes are required to take a server offline or add a new one.

Each site has continuity plans to improve and retire or replace server and storage infrastructure. We use physical machines and virtual servers. We have been evaluating major commercial cloud services by testing our website and software on AWS, Google and IBM cloud services. Currently their cost exceeds the cost of our existing institutional systems, but we may migrate some of our public facing services to the cloud in the near future.

Security is provided by each site's Institutional Information Services group. UniProt data is in the public domain so our security is to protect our servers, institutional networks and production systems (see Security R16 response).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

UniProt takes long-term preservation of its data and annotation very seriously. As a knowlegebase the preservation of original the sequence submissions are less of a concern as most sequences were derived from other repositories (e.g. INSDC databases). What is critical is the value-added curation of protein function which represents years of effort of the scientific community as well as UniProt staff.

UniProt is dependent on core funding from NIH in the United States and EMBL and SERI in Europe (see R3). See (https://www.uniprot.org/help/about) for funding sources. We have successfully maintained funding since 2002 as UniProt and prior to that as individual entities doing protein sequence curation. Our funding and agreement with submitters do require that we make the data public however they do not include specific contractual language for long-term preservation. However long-term preservation it is a fundamental assumption we make in all our activities.

Bitstream preservation: We backup and archive our data in multiple locations on multiple continents as follows.

In the case of a catastrophic loss of funding where all three sites could not continue to work in a reduced capacity, we will archive the data with documentation of their formats, schemas, and annotation descriptions and continue to make them public and available to partner or successor institutions. We expect this to happen via a number of mechanisms.

● The UniProt consortium members EMBL-EBI, SIB and PIR are part of larger institutions that support the project and investigators with additional funds and computational resources. UniProt Websites and FTP sites are mirrored and backed up at these organisations. This would continue at all or some of the sites possibly at reduced level.
● UniProt data is free to reuse and many other informatics resources and repositories redistribute copies of each release via FTP or their own database.
● Some of the large cloud service providers (Amazon, Google) have programs to provide low cost and even free storage and distribution of datasets for research. We are applying to make our data available there.

Ensuring future interpretability of digital assets: Our backups include documentation of their formats, schemas, and annotation descriptions. In case of funding loss there are additional internal process documents we can archive as well, such as our curation manual describing details of how and what we curate. We continue to investigate new formats, databases and other technologies to use as our data increases in size and complexity. Currently we are looking at uses for knowledge graphs (GraphQL) and incorporating Big Data strategies in our production and archiving e.g. Spark and Hadoop.

Please see our answers to R1 and our explicit statement on our website that we are committed to preservation (https://www.uniprot.org/help/about). Also see answers to R3 Continuity of Access above.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

For the next renewal, it would be good to have your preservation practices formalized in a preservation plan in order to achieve Compliance Level 4.

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

UniProtKB is divided into two main sections - reviewed (Swiss-Prot) and unreviewed (TrEMBL). The reviewed section is manually curated by expert biocurators, who are experienced, generally PhD-level, biologists or biochemists with a strong background in wet lab research. The responsibility of these biocurators is to read and assess the appropriate literature on a specific protein, and transfer this into the corresponding entry in the database. Further expertise is leveraged through our interactions with related efforts, such as the model organism databases, and through consultations with domain-specific experts. For example, a set of proteins related to Alzheimer’s Disease has been recently updated in consultation with a number of clinicians and researchers active in the field.

Expert curation in UniProtKB/Swiss-Prot follows a well-defined process to ensure that all records are handled in a consistent manner (see https://www.uniprot.org/docs/sop_manual_curation.pdf for a more detailed description of the process). It includes manual verification of each protein sequence as well as a critical review of experimental data from the literature and predicted data from a range of sequence analysis tools. Publications are read in detail and fully curated. Curators assimilate all the information from various sources, reconcile any conflicting results and compile the data into a concise but comprehensive report, which provides a complete overview of the information available about a particular protein (PMID:21447597, PMID:24622611). Every annotation is independently double-checked prior to integration into the entry.

The unreviewed section of UniProt is automatically annotated by transferring well conserved annotation from characterized members in manually curated families of proteins onto less well studied orthologues, identified by protein signature algorithms. The annotation rules enabling this annotation transfer are again, created and evaluated by expert biocurators (see https://www.uniprot.org/help/automatic_annotation, and http://insideuniprot.blogspot.com/2015_11_01_archive.html). The rules can annotate protein properties such as the protein name, function, catalytic activity, pathway membership, and subcellular location, along with sequence specific information, such as the positions of post-translational modifications and active sites. All predictions are refreshed with each UniProtKB release to ensure the latest state-of-knowledge predictions (https://www.uniprot.org/help/unirule).

All information is linked back to its original source using the Evidence and Conclusion Ontology, a community standard for evidence information, linking each statement to its source evidence and enabling users to evaluate the standard of data.

All curation procedures used throughout the database are documented and made publicly available.
https://www.uniprot.org/help/biocuration
https://www.uniprot.org/help/manual_curation

In addition, UniProt is part of the Gene Ontology and IMEx consortia (PMID:22453911, PMID:30602777), and collaborates with Rhea (PMID:27789701) providing annotations for GO terms, protein-protein interaction data and enzymatic reactions, respectively.

The community can submit corrections and updates to data through the feedback form provided on the UniProt website at https://www.uniprot.org/update. This feedback goes to our helpmail system (help@uniprot.org). This goes in parallel with a regular review of information already present in the database which is part of our standards procedures.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Our overall workflows are shown in this image from our websites. This includes production and annotation processes plus some specialized dataset workflows on complete proteomes and UniRef clusters.

FIGURE 2 (UniProt overall workflow) here

We outlined our import workflows in our response to R8 above. Our unique annotation workflows are described in R11 above, the addition of additional information added to each entry in the UniProt database follows a strictly defined set of rules and procedures. Additional data may be added to entries from external sources, but this is checked and verified to conform to the UniProt format and standards prior to integration.

Unique Accessions (described in R7) are associated to protein entries when created in the TrEMBL section of UniProtKB and are maintained along the different steps of the curation process and throughout any further steps in ‘life’ of a protein entry. We keep a track of the history of an entry in UniSave (UniProtKB Sequence/Annotation Version Archive; http://www.ebi.ac.uk/uniprot/unisave/app/#/) and allow users to check it from release to release. Unlike UniProtKB, which contains only the latest Swiss-Prot and TrEMBL entry versions, UniSave provides access to previous versions of these entries. Archived versions of a UniProtKB entry are accessible through the Previous versions link located at the bottom of the entry view's left-hand navigation bar. (example: www.uniprot.org/uniprot/O14733?version=*)

Archiving of protein sequences is one of the tasks undertaken by the UniProt resource, although as previously described our main focus is on curation, quality checking, tracking evidence and provenance. The UniParc section of the UniProt resource (https://www.uniprot.org/help/uniparc; see also R7 and R8) is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. Proteins may exist in different source databases and in multiple copies in the same database. UniParc avoids such redundancy by storing each unique sequence only once and giving it a stable and unique identifier (UPI).

Additional workflows, shown in the figure and described in the links that follow, include those for creation and distribution of complete proteomes (https://www.uniprot.org/help/proteome), proteome clusters (https://www.ncbi.nlm.nih.gov/pubmed/?term=27153712+21556138) and UniRef clusters

(https://www.ncbi.nlm.nih.gov/pubmed/?term=25398609+17379688).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

An important aim of UniProt is the production of a resource to ensure that our data is freely available. As a community resource, UniProt has been recognized as an exemplar implementation of the FAIR (Findable, Accessible, Interoperable and Reusable) principles (nature.com/articles/sdata201618). Continuous development is needed to ensure delivery of this key component of the life science infrastructure. The UniProt website is used by hundreds of thousands of scientists every month.

Does the repository offer search facilities?
UniProt allows basic and advanced (Boolean) search by text queries through:

- Its website, see (https://www.uniprot.org/help/text-search)
- Its website and Protein REST APIs, a SPARQL API and JAVA API, see (https://www.uniprot.org/help/programmatic_access)

● We also allow sequence similarity searches via BLAST (https://www.uniprot.org/blast/) and peptide matching (https://www.uniprot.org/peptidesearch/)
● Batch retrieval of entries via lists of UniProt accessions or identifiers from other repositories is also possible (https://www.uniprot.org/uploadlists/)

Does the repository maintain a searchable metadata catalogue to appropriate (internationally agreed) standards?
Yes, see:
www.uniprot.org/core/
www.uniprot.org/help/?query=*&fil=section%3Amanual
ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot.xsd

Does the repository facilitate machine harvesting of the metadata?
Yes, via the APIs and metadata catalogues above.

Is the repository included in one or more disciplinary or generic registries of resources?
Three important ones are listed below.
https://fairsharing.org/biodbcore/?q=uniprot
https://datamed.org/search.php?query=uniprot&searchtype=repository
https://www.elixir-europe.org/platforms/data/core-data-resources

Does the repository offer recommended data citations?
To cite the database, see these publications (https://www.uniprot.org/help/publications)
To cite the data, use the entry accession numbers and web links described below.
https://www.uniprot.org/help/accession_numbers
https://www.uniprot.org/help/linking_to_uniprot

Does the repository offer persistent identifiers?
Yes, each record in UniProt has a persistent identifier, which we refer to as an ‘accession number’ (https://www.uniprot.org/help/accession_numbers). We also assign a PURL (Persistent Uniform Resource Locator) to each record (e.g. http://purl.uniprot.org/uniprot/P99999) for use as URIs in our RDF representation. We are maintaining a resolution service for these PURLs (to retrieve either the RDF or the webview of a record) since many years. More recently two projects, identifier.org and N2T.org, teamed up to establish global standards for the citation of biomedical data (doi:10.1038/sdata.2018.29) and UniProt is also registered in this system (https://registry.identifiers.org/registry/uniprot) to provide an alternative route to resolve our records (e.g. https://identifiers.org/uniprot:P99999).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The data in UniProt and the scientific literature has been collected and updated over many decades with careful attention to making it understandable and reusable. Formats, terminologies and processes have changed over that time usually do to: community input; changes in experimental techniques; and, UniProt coping with the ever-increasing capacity to sequences DNA. We participate in many community efforts to support new formats (XML, GFF, RDF), new ontologies and terminologies (UniProt is part of the Gene Ontology consortium) and other standards (Human Proteome Standards, HUGO Human Gene Nomenclature Committee and Human Genome Variation Society guidelines). Answers to specific questions are:

Which metadata are required by the repository when the data are provided (e.g., Dublin Core or content-oriented metadata)?
Content oriented metadata are required for input. For direct protein sequence submission, the metadata requirements are defined here: https://www.ebi.ac.uk/swissprot/Submissions/spin/help. Protein sequences from translations of DNA coding regions come with metadata defined by the INSDC databases like EMBL-ENA which has standards and a metadata model (https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html). At minimum a sequence must have a 1) source organism with taxon ID, 2) an identifiable submitter with mailing address. This can be individuals or an

organisation or a combination. Ideally much more will come including a Protein Name, Gene Symbol, published Citations and other initial annotations on structure or function. We use the Dublin Core vocabulary in our RDF representation where applicable (i.e. for bibliographic references).

UniProt contains well over 100 data types (see https://www.uniprot.org/owl/core.rdf) and there are only few that can be covered with standard metadata schemes like the Dublin Core (for bibliography), or community vocabularies like the Sequence Ontology and FALDO (for the description of sequence ranges). UniProt is actively engaged in the developments of these domain-specific vocabularies to extend them and increase their adoption.

Are data provided in formats used by the Designated Community? Which formats?
Yes, UniProt provides data in multiple formats for different communities and purposes. Main formats are:

FASTA - format of sequence analysis tools
UniProt Flat File - Our legacy format still widely used and supported.
XML
RDF/XML
Tab delimited - customisable on the website.
Excel - customisable on the website.
GFF - General Feature Format (in version 3)

Are measures taken to account for the possible evolution of formats?
Yes, we are involved in the community and aware of new formats that might be useful for out data. Not all the formats listed above were always available and some have undergone modifications. Some APIs and related formats have been retired such as SOAP services and DAS Distributed Annotation Service. We have recently added JSON formats to the Protein API.

Are plans related to future migrations in place?
Nothing definite now, some changes may occur as we refactor our website and associated APIs. We always give advance notice to users of significant changes through the website and email lists.

How does the repository ensure understandability of the data?
We have extensive documentation of our annotation processes, terminology and formats. We have daily communications with users via the help mail. We have help videos, FAQs and blogs on how to best use the data. We conduct usability studies and workshops in the research community.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

Accept

TECHNOLOGY

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The UniProt database are produced by the coordinated efforts of staff and infrastructure at three locations: The European Bioinformatics Institute (EMBL-EBI; https://www.ebi.ac.uk), the SIB Swiss Institute of Bioinformatics (https://www.sib.swiss) and the Protein Information Resource (PIR; https://proteininformationresource.org). The databases are published via an interactive website, APIs and FTP site in different serialization formats, including the W3C standards XML and RDF, as well as life sciences community standards such as FASTA, GFF, PEFF.

All sites run servers with the Linux OS. Teams of system administrators ensure that the servers are updated with the latest security patches and perform updates to new OS versions. The resource also used both commercial (ORACLE) and open source (Postgres) RDBMS that are managed by a team of DBAs that support the software development teams.

Public services are hosted at multiple locations to provide resilient services.

The data growth as well as the usage of the public services are monitored closely to periodically evaluate the need for new hardware (compute, storage, network). The growth in compute power broadly aligns with the growth in storage

requirements. As biological instruments increase so do networking, storage and compute power to process and store information. In addition to compute expansion in relation to storage, computational requirements may also increase due to new or advanced processing techniques or services.

The resource has software developers at the three sites that implement software that is tailored to the needs of the resource and its users. This software (typically, but not exclusively, Java applications) relies on many open source libraries and frameworks. The code is managed with a revision control system (Git) and issue tracker (Jira). Internal documentation is maintained in Confluence.

The resource monitors the uptime and performance of its public services to ensure that the global responsibilities of the resource are adequately met by the connectivity to public networks and bandwidth.

EMBL-EBI's state-of-the-art technical architecture is secure and robust, and is distributed in three discrete data centers in different geographical locations to assure long-term security. This gives our data very high protection through redundancy, and provides sufficient capacity and reserve to ensure our management of the rising influx of data and compute requests. As of spring 2018 the main compute farm has 34,000 cores (27,000 high throughput and 7,000 high performance) and the installed disk-based storage capacity is above 200 Petabytes. The internal network has a 100 Gigabit backbone within its data centers and multiple 10 Gigabit connections between data centers and most servers are in these data centers are connected by at least 10 Gigabit networks. EMBL-EBI has two independent 10 Gigabit physical uplink from the data centers to Janet, Internet2 and Geant (the UK, pan-American and pan-European research networks, respectively). EMBL-EBI has contracted with Kao Data services in the UK to deliver data center services.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

As a globally distributed knowledgebase we adhere to multiple IT security plans from our institutions plus multiple legal frameworks. As all our data is in the public domain our main concern is in protecting our servers and institutional networks from malicious activity. As for user privacy we are following the new European Union General Data Protection Regulation (GDPR) rules and practices. Our data has a strict internal Electronic Data Security Policy (EDSP), the implementation of which is overseen by the IT Security Steering Committee in our institutions.

All three site maintain physical and electronic security processes and systems. Data backup processes were answered in R9 above.

Physical security of the Data Centers is provided by locked doors, private server cages including mesh walls, dedicated cameras and some cases biometric hand scanners.

Electronic security, all systems use Virtual Private Networks (VPN) and an institutionally managed password and dual identification systems to access the networks and data systems. All systems are protected by network firewalls and security protocols. Server and network access are monitored 24*7 by Institutional Information security offices or private security contractors using a variety of management tools.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

We found this application somewhat confusing in part because it was designed for strict repositories and not a knowledgebase like UniProt which is supposed to do more than preserve the data. So, it was a little more difficult to know how best to answer some questions.

Also, some of the guidance seemed redundant as questions on security, continuity and other processes appeared in several places making our answers somewhat redundant. We tried to fix this, but some redundancy was needed to provide the best answers. In part our unique history and consortium structure made some answers complicated as infrastructure varies across the different institutions.

We had some figures in the application and the online system only takes text so had to upload entire document. Hope this is not a problem but figures help.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: