DELIVERABLE 10.2

DATA MANAGEMENT PLAN

WORK PACKAGE NUMBER: WP10

WORK PACKAGE TITLE: PROJECT MANAGEMENT

TYPE: REPORT

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Document Information

iReceptor Plus Action Information	
Action full title	Architecture and Tools for the Query of Antibody and T-cell Receptor Sequencing Data Repositories for Enabling Improved Personalized Medicine and Immunotherapy
Action acronym	iReceptor Plus
Grant agreement number	825821
Project coordinator	Prof. Gur Yaari
Project start date and duration	1 January 2019, 48 months
Project website	https://www.ireceptor-plus.com

Deliverable Information	
Work package number	WP10
Work package title	Project Management
Deliverable number	D10.2
Deliverable title	Data Management Plan
Description	
Lead beneficiary	Bar Ilan University (BIU)
Lead Author(s)	Jos Dumortier (time.lex), Liesa Boghaert (time.lex)
Contributor(s)	Jos Dumortier (time.lex), Liesa Boghaert (time.lex)
Revision number	
Revision Date	

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Status (Final (F), Draft (D), Revised Draft (RV))	D
Dissemination level (Public (PU), Restricted to other program participants (PP), Restricted to a group specified by the consortium (RE), Confidential for consortium members only (CO))	CO (including the Commission Services)

Document History

Revision	Date	Modification	Author
1	09.06.2019	Initial version	Jos Dumortier and Liesa Boghaert
1	10.06.2019	review	Gur Yaari
1	14.06.2019	Comments and questions	Artur Rocha
1	15.06.2019	review	Felix Breden
1	19.06.2019	Comments	Artur Rocha
2	24.06.2019	edit	Jos Dumortier
2	26.06.2019	edit	Encarnita Marriotti-Ferrandiz
3	28.06.2019	Final version	Jos Dumortier

Approvals

	Name	Organisation	Date	Signature (initials)
Coordinator	Prof. Gur Yaari	Bar Ilan University	30.06.2019	GY
WP Leaders	Prof. Gur Yaari	Bar Ilan University	30.06.2019	GY

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

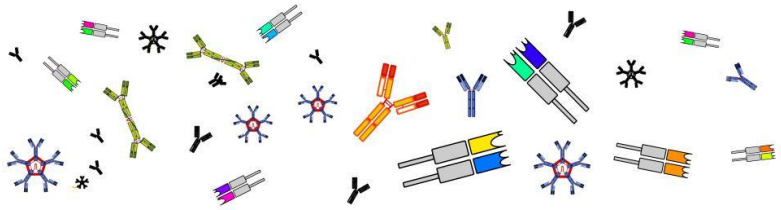


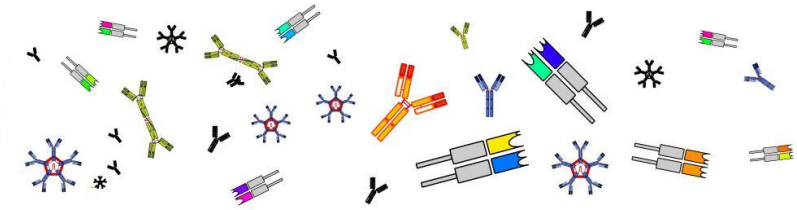
Table of Contents

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Abbreviations

DMP	Data Management Plan
FAIR	Findable, Accessible, Interoperable, Re-usable
EC	European Commission
AIRR-seq data	Adaptive Immune Receptor Repertoire sequencing data
iReceptor Plus Platform	New scalable platform integrating immense AIRR-seq data from multiple repositories, with new analysis tools and linking to other omics type data sets
MiAIRR standard	AIRR Community-endorsed standard that describes the 'Minimal Information' for repertoire metadata and sequence annotation data in studies that utilize AIRR-seq data
AIRR Data Commons	Distributed network of AIRR-seq data repositories that adhere to the AIRR Community standards
AIRR Data Representations	AIRR Community-endorsed file format specifications for storing large amounts of annotated AIRR data

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



Executive Summary

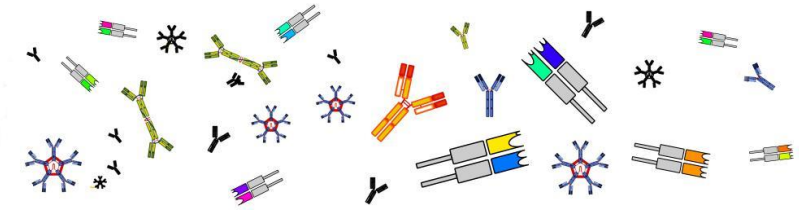
The purpose of this deliverable is to set out the Data Management Plan (DMP) for the iReceptor Plus action.

This document should be regarded as a living document, that will be updated throughout the action.

It provides information on:

- What categories of research data will be collected, processed and/or generated by the action;
- Which methods and standards will be applied in respect of the FAIR data principles for handling research data during and after the end of the action;
- Whether research data will be shared or made publicly accessible and if so, following which open access model;
- How data will be curated and preserved, during and after the end of the action;
- Which additional safeguards will be implemented to ensure respect of the FAIR data principles in terms of allocation of resources, data security, research ethics and intellectual property rights.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



1 Introduction

Data Management Plans (DMPs) are considered to be a key element to sound data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 action.

The present document constitutes the first version of the iReceptor Plus action DMP and reflects the status after the first six months of the action. It is based on the DMP template provided by the European Commission (EC)[1] and follows the 'Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020'[2].

This DMP should be regarded as a living document that will evolve and gain more precision and substance during the project implementation. Hence, information will be made available on a finer level of granularity through updates as the implementation of the action progresses and when significant changes occur, such as (but not limited to) the inclusion of new data, changes in consortium policies (e.g. new innovation potential, decision to file for a patent…), changes in consortium position and external factors (e.g. new consortium members joining or old members leaving).

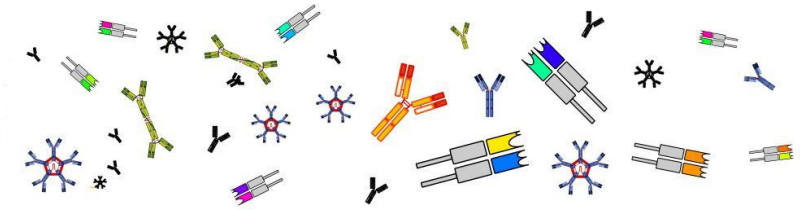
As a part of making research data **Findable, Accessible, Interoperable** and **Re-usable (FAIR)**, this DMP includes information on:

- What categories of research data will be collected, processed and/or generated by the action;
- Which methods and standards will be applied in respect of the FAIR data principles for handling research data during and after the end of the action;
- Whether research data will be shared or made publicly accessible and if so, following which open access model;
- How data will be curated and preserved, during and after the end of the action;
- Which additional safeguards will be implemented to ensure respect of the FAIR data principles in terms of allocation of resources, data security, research ethics and intellectual property rights.

[1] See European Commission, Guidelines on FAIR Data Management in Horizon 2020, version 3.0, 26 July 2016.

[2] European Commission, Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



2 Data Summary

2.1 Research data

The notion 'research data' refers to "information, in particular facts or numbers, collected to examined and considered as a basis for reasoning, discussion or calculation".[3]

Examples of research data include (among others) statistics, results of experiments, measurements, database contents, protein or genetic sequences, survey results, contents of applications, interview recordings and images. Trade secrets, commercially sensitive information and confidential information are however not considered to be research data.

Given the context of the EC Open Research Data Pilot, this DMP focuses primarily on research data that are available in digital form.

2.2 Collection purposes

Research data will be collected[4] and processed in the course of the action for the following purposes and in relation to the following project objectives:

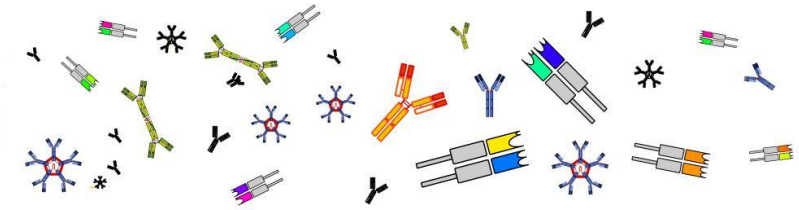
- **PROMOTE** human immunological data storage, integration and regulatory compliant sharing[5] for a wide range of clinical and scientific purposes;
- **DEVELOP** an innovative platform to integrate distributed repositories of Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) data;
- **ENABLE** improved personalized medicine and immunotherapy in cancer, inflammatory and autoimmune diseases, allergies and infectious diseases;
- **LOWER** the barrier to share, access and analyse large AIRR-seq data sets from around the world and **EASE** the availability of these AIRR-seq data to academia, industry and clinical partners;
- **ADVANCE** the understanding of immune responses and provide new targets for therapies and new methods for monitoring therapeutic efficacy;

[3] European Commission, Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017,4.

[4] Note that iReceptor Plus will create a network of repositories and will thus only *bring together* data that already exist in the distributed repositories. It will not in itself collect research data from individuals but will merely facilitate access to pre-existing data originating from the repositories that are part of the iReceptor Plus network.

[5] Regulatory compliant data sharing will be monitored and ensured through several deliverables, such as D3.3, D11.2 and D11.12.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



- **PROMOTE** the discovery of biomedical interventions that manipulate the adaptive immune system such as vaccines and other immunotherapies.

In doing so, the iReceptor Plus action will benefit (1) the design of drugs and vaccines, (2) the development of new therapeutic approaches and (3) the identification of new biomarkers for diagnosis and prognosis.

2.3 Data sets

In the first months of the action, the relevant action partners were asked to describe the data sets that will be processed in the course of the iReceptor Plus action.

At this early stage of the action, four fundamental categories were identified:

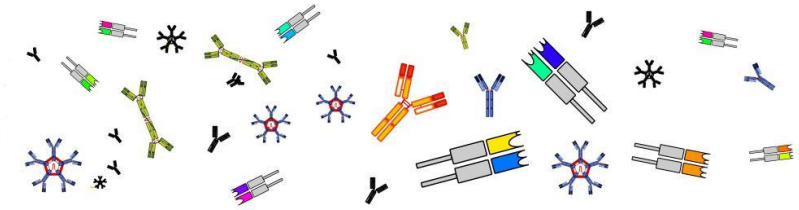
- **Study Metadata:** These research data concern the metadata relating to a given research study that is included in a repository that is part of the iReceptor Plus network of distributed repositories. These data typically consist of data about the study, the subjects in the study, the biological samples in the study as well as the way in which these samples were obtained and processed.

- **AIRR-seq data:** These research data are the sequences that result from the processing of biological samples from research studies, annotated with a set of features critical to understand the Adaptive Immune Response.

- **External Metadata:** These research data encompass metadata that are linked to the 'core' data components above but originate from outside the iReceptor Plus Platform. Examples of External Metadata linked to Study Metadata are other immunologically relevant data such as cell phenotype, microbiome, clinical, gene expression, HLA-typing, quantified autoantibodies, cytokine and chemokine data. Genetic associated data will be provided as quantitative measures without public access to the raw sequence (except in regards of point 2.4). External Metadata linked to AIRR-seq data are for example epitopes (binding target).

- **Analysis Metadata:** These research data are generated from analysing the combination of a set of Study Metadata, AIRR-seq data and External Metadata and as such constitute 'data about other data'.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



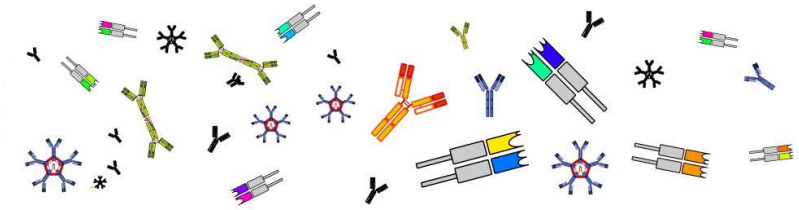
Apart from the four categories mentioned above, some other types of data might be collected, processed and/or generated in the course of the action:

- Documents relating to meetings, focus groups, workshops…
- Academic literature and knowledge materials

All partners commit to continuously keep track of the specific data sets processed under the tasks they are leading and to report them internally by completing the following table:

[**PARTNER INITIALS**], **WP** [No.], **T** [No.] (e.g. time.lex, WP3, T3.4)	
Data set	*Name of the data set and reference*
Description	*Description, source of the data, creation method…*
File format	*Software or file format used to work with the data – e.g. Word, Excel…*
Metadata	*Data characteristics*
Data sharing	*Data derives from…, is shared with…, is used by…*
Archiving and reservation	*Data are stored… Backups…*
Additional information	*Are you generating the data or sourcing it from elsewhere? Are there certain terms and conditions applicable?* *How will the data be created or collected? What instruments or tools will be used?* *What transformations will the data undergo? What software or file formats will you use as you work with the data?* *Will the data be updated or become redundant as you make revisions and produce subsequent versions?* *Are you processing information that falls outside the scope of this DMP? (e.g. sensitive or confidential information)* *Is there ethics approval or is ethics approval required?*

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



2.4 Origin of data and re-use of existing data

The objective of the iReceptor Plus Action is to build a common scalable platform to integrate distributed repositories of AIRR-seq data for enabling improved personalized medicine and immunotherapy for diseases with an immune component. iReceptor Plus will be designed as a network of distributed repositories that facilitates data queries and advances analyses through one or more web portals. As such, the iReceptor Plus platform will enable the search and re-use of existing data, originating from the AIRR-seq data repositories of individual labs, research institutions etc. that each maintain control over their data and ascertain compliance with their local legislation through their own data stewards.

However, to achieve this end goal, in the course of the action, the iReceptor Plus partners will also base their research on data that does not originate from the distributed repositories that are part of the action. For example, when developing software to establish the iReceptor Plus platform, end-users and stakeholders will be consulted to consider users' requirements in the development phase. This will entail the collection of new data by the project partners. Furthermore, existing knowledge materials will be consulted to support the performed research. These existing materials will be selected on the basis of their relevance to the project objectives and will be collected following scientific collection methods, using databases and repositories available to the project partners.

2.5 Expected size of the data

The total size of the research data is difficult to estimate at this point in the action. This is because the size of the data will be largely dependent upon the number of AIRR-seq data repositories that will participate in the AIRR Data Commons repository network.

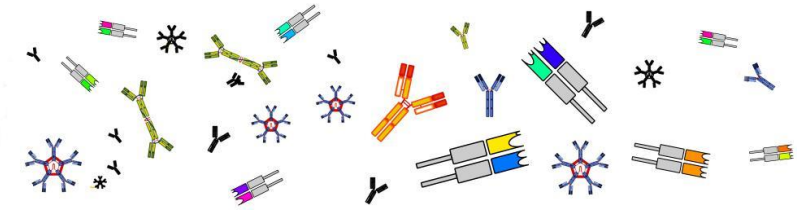
However, the objective of iReceptor Plus is to create '*a critical mass of approximately 10 well curated, international repositories within the AIRR Data Commons each with 100s of millions or billions of annotated sequences, that are of scientific significance to the general AIRR Community.'* This would allow iReceptor Plus to "scale-out" the AIRR Data Commons repository network to tens or hundreds of independent repositories at the end of the action.

Apart from the annotated sequence data, the iReceptor Plus action will also process metadata relating to the research studies underpinning these sequence data, external metadata (as described above)[6] and data relating to knowledge materials. The total size of the research data is therefore also dependent upon the number of studies incorporated in the repositories and the number of data per study. Moreover, iReceptor Plus will generate data that stem from analysing the combination of data from different repositories. The amount of data generated from this analysis is hard to predict at this stage.

In general, however, it can be assumed that the expected size of the data to be collected, processed and generated in the course of iReceptor Plus will be rather large. It is estimated to be at a maximum 100 terabyte.

[6] See section 2.3, p. 9.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

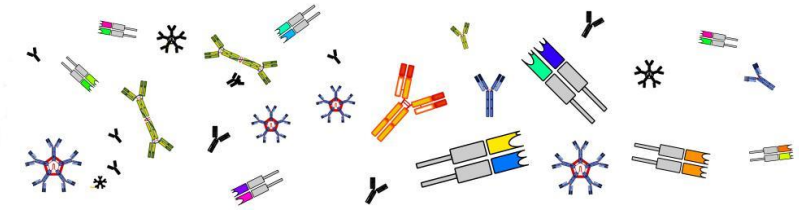


2.6 Data utility

The iReceptor Plus action aims to lower the barrier to share, access and analyse large AIRR-seq data sets from around the world and aspires to ease the availability of these AIRR-seq data to academia, industry and clinical partners.

Therefore, the results of the action research activities and platform established as a result of the action will primarily be of benefit to academic researchers, clinicians and industrial partners. Nevertheless, from a more holistic point of view, the data processing and analysis performed by means of the iReceptor Plus platform will contribute to the well-being of the society as a whole through the subsequent design of drugs and vaccines, the development of new therapeutic approaches and the identification of new biomarkers for diagnosis and prognosis.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



3 Findable, Accessible, Interoperable and Re-usable data (FAIR data)

3.1 Making data findable, including provisions for metadata

The iReceptor Plus action attaches great importance to making its research data findable, discoverable and identifiable. That is why the action will build on the achievements of the present iReceptor platform developed at Simon Fraser University, the principal Canadian partner in the iReceptor Plus Consortium. In the context of iReceptor, the beneficiaries of the action contributed to the development by the AIRR Community of a minimal standard for repertoire metadata and sequence annotation data for studies involving AIRR-seq data (the MiAIRR standard). This standard lists the minimum data elements to be included with an AIRR study for publication. In addition, a file format standard for storing and sharing an extended set of sequence annotation data targeted at data exchange between advanced AIRR-seq analysis tools was defined.

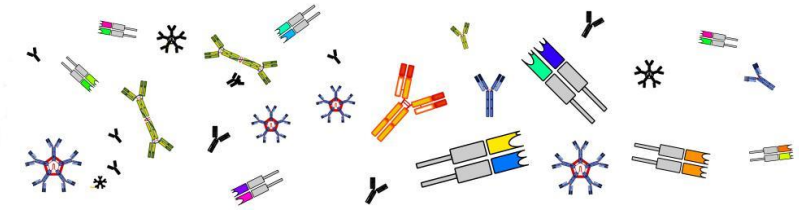
By endorsing these standardisation techniques, iReceptor Plus increases the findability of AIRR-seq data and Study Metadata. Moreover, these standardisation efforts allow researchers skilled in the art of AIRR-seq data generation and analysis to reproduce results of a given study. As such, they provide the foundation for a unified environment of data sets and analysis tools, thereby creating economies of scale for individual researchers. Although no internationally accepted standards for re-use of External Metadata and Analysis Metadata (as defined above[7]) exist, iReceptor Plus aims to also make these data Findable, Accessible, Interoperable and Re-usable (FAIR) to the extent possible.

As a consequence, most research data will be discoverable with metadata, while some data may even be identifiable and locatable by a portable unique identifier (PUID). In general, however, there will not be introduced globally unique identifiers for the data. Nevertheless, each individual repository will ensure its identifiers are unique within the repository in a way that the combination of the repository URL and the identifier will uniquely identify the digital object. Moreover, ontologies for field values will be developed to enhance findability and interoperability. Finally, search keywords will enrich the annotation of the data, which will facilitate and optimize its re-use.

For project internal sharing, such as the sharing of unfinished reports and meeting minutes, iReceptor Plus uses ‘ownCloud’ (a cloud storage software suite) in order to efficiently manage the project information amongst the project partners and to enable the preservation and appropriate versioning of documents. Moreover, the partners commit to using appropriate naming conventions as defined by the coordinator and agree to ensure clear versioning by indicating ‘v1’, ‘v2’, … or ‘final’ at the end of the name of a document whenever multiple versions of a document exist.

[7] See title 2.3.

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



3.2 Making data openly accessible

The iReceptor Plus platform is fundamentally a data sharing platform with open access as one of its foundations. It will be introduced as free and open license software, making it possible for the research community to extend and adapt tools and technologies utilized within the action, e.g. for research labs to integrate their own data repositories in the platform.

However, many biopharmaceutical companies, producing some of the most significant AIRR-seq data cannot integrate their data into any implementation of an AIRR-seq Data Commons unless they can protect their data in terms of security and licensing. For this reason, iReceptor Plus, unlike other attempts at AIRR-seq integration, will add layered data security and will establish three levels of access/security.

The first level of the platform (1) will allow access to completely open data in the AIRR Data Commons. On a second level (2), iReceptor Plus will provide the ability to manage private, protected data that the researcher wants to compare to data in the AIRR Data Commons, without exposing their data. Finally (3), the iReceptor Plus platform will provide an intermediate level of sharing, e.g. among consortia of repositories that share common consent structure or reciprocal data transfer agreements (DTA). This way, the platform will not only provide access to large amounts of open data but will also facilitate the comparison of protected data with open access data.

In terms of software, access to the data will be dependent on the installation of the AIRR Data Commons API, which operates as the primary interface between a data repository and the outside world. The API provides programmatic access to query and download AIRR-seq data. It uses JSON as its communication format, and standard HTTP methods like GET and POST.

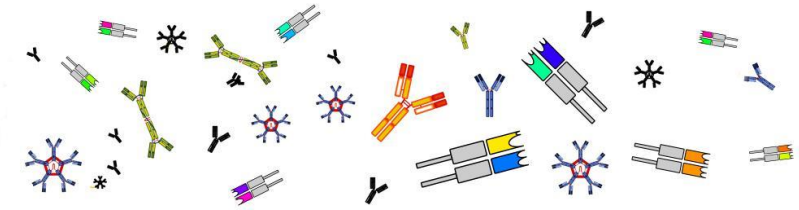
Research data will remain deposited in the distributed repositories, while code and documentation will be deposited in the iReceptor Plus GitHub[8].

Furthermore, within the limits of intellectual property protection and in line with the Horizon 2020 Open Access policy[9], iReceptor Plus will disseminate the achieved results of the research through the publication of public project deliverables on the open section of the website. Similarly, they will be deposited in other existing repositories of the European research results such as Zenodo or OpenAIRE, in accordance with the Horizon 2020 Open Access policy.

[8] GitHub is a leading software development platform.

[9] See the Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



3.3 Making data interoperable

As mentioned above, iReceptor Plus will adhere to the AIRR Community protocols and metadata standards. This will render the iReceptor Plus Platform interoperable with other efforts in this area.

Also, iReceptor Plus will adopt the 'AIRR Data Representations', file format specifications for importing and exporting large amounts of annotated AIRR data. These are versioned specifications that consist of a file format and a well-defined schema. The schema is provided in a machine-readable YAML document that follows the OpenAPI v2.0 specification. The schema defines the data model, field names, data types, and encodings for AIRR standard objects. Strict typing enables interoperability and data sharing between different AIRR-seq analysis tools and repositories, and some fields use a controlled vocabulary or an ontology for value restriction. Specification extensions are utilized to define AIRR-specific attributes.

Lastly, interoperability is increased through the development of an 'iReceptor Turnkey Repository' which will be an open license, freely available software stack that will allow research labs and industry to download, install and securely manage their own AIRR Data Repository as a component part in the AIRR Data Commons. This way, researchers are able to both curate their own original data and render it interoperable with the data already accessible via the iReceptor Plus platform.

3.4 Increase data re-use (through clarifying licenses)

Specific details on licensing will be provided in the deliverable that holds the exploitation plan.

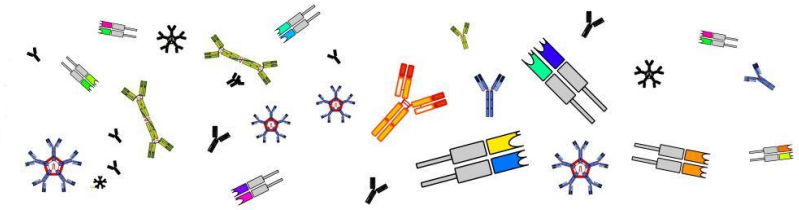
Research data will remain re-usable forever.

4 Allocation of resources

As described in the Description of Action, within the limits of intellectual property protection and in line with the Open Research Access Pilot, iReceptor Plus will disseminate the achieved results through the publication on the open section of the website of public project deliverables. Whether to use Green OA or Gold OA will be decided specifically for each research result.

As defined by the Grant Agreement, the entity responsible for data management is the project coordinator.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



5 Data security

To ensure security of data, iReceptor Plus will maintain an approach of layered data security between the components of the platform, providing multiple levels of authentication, authorisation and auditing. These security layers will provide mechanisms for data stewards to implement control at different levels of granularity. Access may be restricted at multiple levels (e.g. through authentication) to specific types of data (e.g. through role-based authorization) and according to its privacy level. Furthermore, adequate monitoring and auditing mechanisms shall be provided for the previously identified access layers (e.g. through logging and audit controls, including the use of blockchain).

In any case, the consortium partners confirm to comply with the following guidelines in order to ensure the security of data:

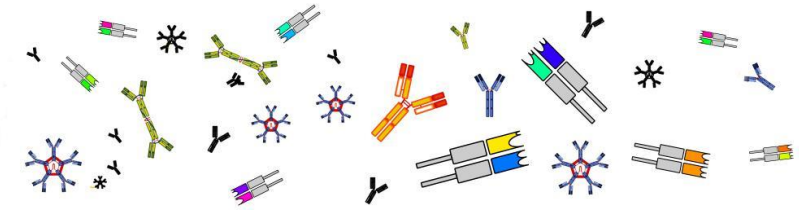
- Encrypt data if it is deemed necessary by the responsible project partner;
- Store data in at least two separate locations to avoid loss of data;
- Limit the use of USB flash drives and personal folders;
- Save digital files in a commonly used format;
- Label files in a systematically structured way in order to ensure coherence.

6 Ethical aspects and intellectual property rights

The ethical aspects of the iReceptor Plus Action will be assessed under Work Package 11, which sets out the ethics requirements that the action must comply with. More specifically, deliverable 11.10 will evaluate the ethics risks related to the data processing activities of the action.

Issues of intellectual property rights and ownership are governed by the provisions of the Consortium Agreement and the Grant agreement, as signed by all project partners.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



7 Conclusion

The purpose of this document is to set out a first, provisional version of the DMP for the iReceptor Plus action. This DMP will be revised and updated throughout the entire duration of the action.

It provides initial information on:

- What categories of research data will be collected, processed and/or generated by the action;
- Which methods and standards will be applied in respect of the FAIR data principles for handling research data during and after the end of the action;
- Whether research data will be shared or made publicly accessible and if so, following which open access model;
- How data will be curated and preserved, during and after the end of the action;
- Which additional safeguards will be implemented to ensure respect of the FAIR data principles in terms of allocation of resources, data security, research ethics and intellectual property rights.

This information will be further updated throughout the action.

8 References

Literature

European Commission, Guidelines on FAIR Data Management in Horizon 2020, version 3.0, 26 July 2016.

European Commission, Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017.

Websites

https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

https://www.ireceptor-plus.com/

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)