Deliverable

D4.5 Metadata catalog operational, with initial content	
Version | Status	V1 | final
Work package	WP4
Lead beneficiary	ULEIC (Anthony Brookes)
Due date	30.09.2018 (M9)
Date of preparation	12.11.2018 (M11)
Target Dissemination Level	Public
Author(s)	Anthony Brookes (ULEIC), Spencer Gibson (ULEIC)
Reviewed by	Ana Rath (INSERM-Orphanet), Sergi Beltran Agulló (CNAG-CRG), Morris Swertz (UMCG)
Approved by	Holm Graessner (EKUT)

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257.

Explanation according to GA Annex I:

Metadata aspects of the project.

Abstract:

In order for the resources, that are contributed to and will be developed during the Solve-RD project, to be utilised to their full potential a discovery system was required. Such a system is being developed based on proven technologies and a suitable set of standards. This deliverable focuses on the building of an initial version of the system, based on the Café Variome platform, for asset discovery called RD-NEXUS (Rare Disease Network for EXploring the UNseen). In order to build the system, a working data model for an agreed set of parameters (termed 'findable facets') was defined and integrated in to a data model. APIs allowing interoperability with other systems were also developed in collaboration with the GA4GH. Exemplar data have been processed and entered into the current RD-NEXUS system to illustrate its functionality and highlight any potential issues, before being demonstrated to potential users within the ERN networks. A complete first version of RD-NEXUS was thereby created, and is now available for demonstration and testing. It can be reached at rd.discovery-nexus.org [user: rdnexus@cafevariome.org password: solverd].

Introduction:

Discovery Services and RD3

The Solve-RD project will encompass many datasets, patients and biosamples which will be useful within and beyond the project. As such, the project aims to professionally track what assets exist, their characteristics, their location(s), their availability, their consents/conditions of use, and methods for requesting and accessing them. This will utilise 'metadata' (contextual data about data and assets), located in a Solve-RD "Rare Disease Data Database (RD3)". Additionally, the project will provide a way for researchers to search for ('discover') the existence of these assets, without directly accessing or revealing information about them. This will be achieved by creating a dedicated 'discovery' layer across the project, which will enable registered, approved users to undertake discovery queries based on the RD3 content and also on some actual data elements.

The spectrum and number of assets that will be tracked via RD3 and made discoverable is large and complex, and will continually grow with time. Furthermore, the European Joint Project in Rare Disease (EJPRD) infrastructure project that will launch in 2019 will have its own focus on RD asset discovery. Therefore, we will build the Solve-RD RD3 and discovery platform in a manner that will allow them to be expanded in liaison with EJPRD to also serve the wider RD community. This is consistent with Solve-RD's desire to support the wider ERN community, not least via enabling them all to advertise, discover and request assets within Solve-RD. This will be accomplished by basing our solutions on interoperability standards (which already exist or are being devised by international collaborations in which we are active) and on a federated 'lattice' approach to asset discovery. Using federated IT architectures will enable different parties to add to and update information about extant assets they are responsible for, and simultaneously control (via carefully managed user lists and permissions) who can query for what items and what response types those users may get. Potential users may be within and/or outside the list of Solve-RD partners.

To design and build this system WP4 is starting with existing proven software solutions and standards, and then adapting these to bring interoperability with other current and future discovery systems relevant to the RD-related field, not least those now emerging from BMMRI,

ELIXIR, GA4GH, RD-Connect, IRDiRC and industry (such as Illumina's 'CaseLog' platform installed in Genomics England Limited and elsewhere). This will maximise the impact of Solve-RD and promote the involvement of any and all ERNs and RD researchers who may have additional RD assets of potential interest, even if not yet deposited into Solve-RD's data environments. It will also enable us to connect with and add extensions to existing global RD patient matchmaking services, which is one of Solve-RD's main discovery-related goals.

Contemporary Discovery Science

Data discovery is not merely a sub-category of data sharing - even though both discovery and sharing activities usually begin with a data 'query' step. This is because the philosophy and approaches behind these respective queries are very different, in that sharing queries need to be highly granular and ideally have zero false positive and false negative response rates (as one is searching for specific items to then access and use), whereas discovery queries can tolerate a degree of imprecision and may tolerate or even benefit from some level of false positives (as the objective is to merely locate assets of likely interest whilst fully protecting the characteristics of those assets and the identity of the human subjects they describe). As such, data discovery (and its methods and security) has become something of a science in and of itself.

To illustrate this point, it can be noted that technologies and software now exist by which discovery can be enabled whilst ensuring that the employed systems:

- can avoid data sharing and subject identification
- can avoid enabling data analysis
- may employ an anonymous, simplified, obfuscated version of the data
- can tailor the derivative data to the specific use cases intended
- may locate these 'queryable data' separately from the original full data
- may employ graphics for querying and/or response actions
- keep response types limited (e.g., to yes/no, counts, hand-off to request interfaces, contact details of asset owners)
- restrict system access to approved, registered users who must login with a secure password or other single- or multi-factor authentication regimes
- restrict permissions of users in terms of what assets their queries may interrogate, and what response types they may receive
- typically log all activities, and potentially apply some monitoring for suspicious activity patterns, even involving Artificial Intelligence approaches

Report:

Café Variome Discovery Platform

Solve-RD will ultimately make use of various established and validated software components and technologies in constructing the RD3 and discovery services for the project. For example, we will use relevant capabilities of the GPAP platform, the MOLGENIS databasing infrastructure, the EGA geno/pheno archive system, EBI's EMBASSY cloud, the RD-Connect biosample catalogue, services being created by JRC and Orphanet, and so on. But this particular Deliverable is concerned solely with the creation of a first version data and patient discovery service, not the wider RD3 asset tracking capability. We call this service the "RD-NEXUS" (Rare Disease Network for EXploring the UnSeen).

To establish a first version of RD-NEXUS we have adapted and deployed the 'Cafe Variome' technology - which is a modular, highly flexible and fully-functional software stack dedicated to addressing asset discovery challenges. This has been developed over almost a decade and applied in various domains, not least identifying clinical trial subjects from federated cohorts, enabling a confidential big-pharma toxicity data network, management of primary care patient records, advertising cardiovascular disease mutation frequencies in the Netherlands, and consortia approaches to rare disease patient comparisons (Ehlers Danlos Syndrome and mtDNA disorders). Cafe Variome is a PHP/Apache based application that utilises a flexible 'Entity-Attribute-Value (EAV)' data model (an extension of the i2b2 star schema) plus integrated Elasticsearch, SQL and BCFtools technologies for efficient faceted searching. Each installation can act as a central system or part of any number of secure networks, with user profiles and permissions managed via an associated Authentication Authorisation Interface (AAI) service.

With RD-NEXUS now set up at a single location (as described below), we can proceed to install the same platform at various partner sites to explore federation approaches (work already underway), and develop policies and Application Programming Interface (API) translators to enable other platforms to join with and expand this network (e.g., GPAP, MOLGENIS, RD-connect sample catalogue, Illumina's CaseLog, the GEL ARK platform, members of the MME network).

Relevant Standards

Two core standards had to be settled on to create the first version of RD-NEXUS: The metadata plus data elements used for querying (together called 'Findable Facets'), and the Application Programmers Interface (API) by which queries entered into query builder interfaces are packaged and communicated to the local or federated search engines.

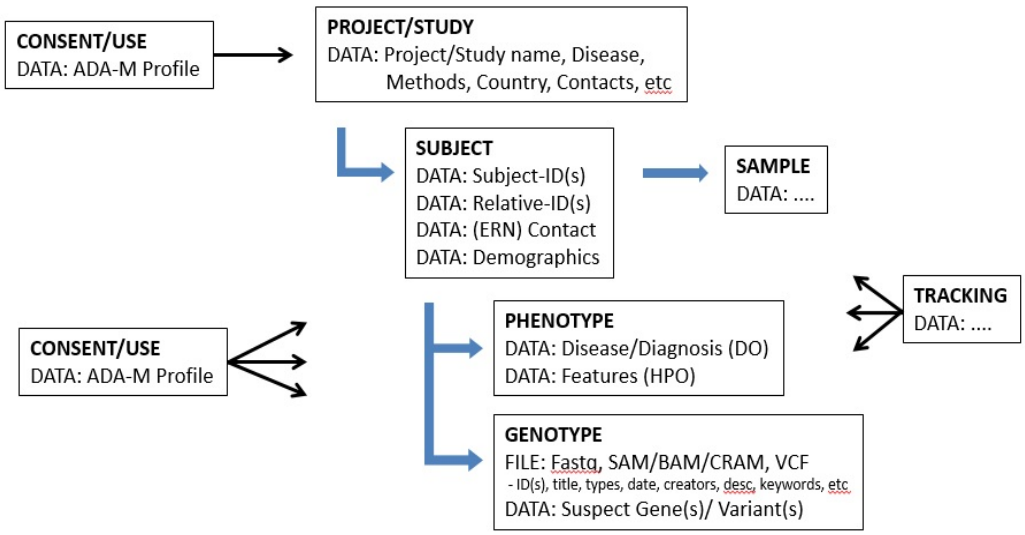

Figure 1: Outline Concept Model for RD3

Via a series of emails, phone discussions, and a face-to-face workshop, over a period of approximately 6 months WP4 partners considered all the mainstream metadata standards and all the data models of Solve-RD project partners, to work out what we would need in terms of Findable Facets for RD-NEXUS. These discussions resulted in a v1 concept model and database model for RD3 and discovery purposes (see Figure 1 and Appendix 1 respectively) and a detailed cross-mapping of all relevant metadata model standards which continues to evolve as we learn more (available at: https://docs.google.com/spreadsheets/d/1FeY-DaKMV3Bz6RbTYi2lOvH9CJESjz6L-KK-bYdNnRk/edit?usp=sharing). The work of defining

these standard fully involved representatives from ERNs ITHACA, GENTURIS and NMD, as the use cases and required functionalities for this IT work must be led by the intended users of the system. Furthermore, it is important to note that any decisions made and implemented at this stage are not assumed to be final. Instead, the aim is to create a first version implementation of RD-NEXUS to present to and discuss with many ERN representatives and other Solve-RD partners, so that everyone can understand the overall objectives and guide WP4 in how to improve the platform so that it provides good real-world utility for ERNs and Solve-RD.

The API employed in RD-NEXUS was developed by Solve-RD partners working in conjunction with and as part of the GA4GH 'Search API' team. This group took about a year to settle on a robust API framework, which itself was closely based upon the original API of the Cafe Variome system. The GA4GH Search API (latest version available here: https://github.com/ga4gh-discovery/ga4gh-discovery-search) accommodates only a few Findable Facets, and has no ability to define operators - and as such it needs further development before it can be approved by GA4GH as a formal v1 standard. Since that may take some time, WP4 is in the process of forking the standard to add in many more Findable Facet components and a full range of query Operators and thereby generated a sufficiently powerful API for use in RD-NEXUS. Our work on this front will be fed back to GA4GH for their consideration.

First version RD-Nexus

Using the above RD3 concept/data model and the extended GA4GH API, we were able to quickly (<1 week) build these into a copy of the Cafe Variome software to constitute RD-NEXUS v1. This is presently hosted by the University of Leicester project partner.

Exemplar data were gathered, so that we could work out how to suitably process/transform discoverable datasets, populate them into the RD-NEXUS discovery platform, and hence have a complete setup to demonstrate to potential adopters. We did not seek to incorporate real patient data in this initial implementation, as it is necessary to first ensure everything works as intended. Now that goal has been reached we will worked with ERNs to progressively populate the system with whatever Findable Facets they would wish to include, for whichever cases they wish to make discoverable, with them being in control of how and by whom this information is to be searched.

We have incorporated three mock datasets: (i) a set of 15 VCF files including variant information (including causative mutations) derived from Whole Exome Sequencing (WES), plus accompanying patient phenotypes (provided by GPAP); (ii) several tens of records from biosamples (provided by the RD-Connect sample catalogue), including basic demographic, disease and phenotype data; and (iii) collections of suspect variants for each of several tens of RD patients, including HPO phenotype information. Custom scripts were generated to convert these into the required EAV format, and these were then imported and indexed within the RD-NEXUS tool, ready for querying. All the imported data fields were consistent with the data fields we currently listed within the RD3 concept and data model.

A complete first version of RD-NEXUS was thereby created, and is now available for demonstration and testing. RD-NEXUS domains have been created at 'rd.discovery-nexus.org' and 'rd.discoverynexus.org' and we have implemented a multi-user account so that people can test-drive the platform:

User: rdnexus@cafevariome.org
Password: solverd

Figures 2-4 present screenshots of the query builder interface, illustrating how queries are entered and how the results of those queries are displayed.

Figure 2: Main query builder interface for RD-NEXUS. The query builder is designed to allow rapid construction of precise queries by presenting fields and values for the discoverable data in a straightforward interface.

Figure 3. Findable Facet Selection. Fields present within each queryable category are updated live from the available data across RD-NEXUS preventing wasted searches that would yield no information. Available fields within the "Patient Characteristics" and "Phenotype Detail" categories are displayed and selectable, to simplify the process of building a discovery query.

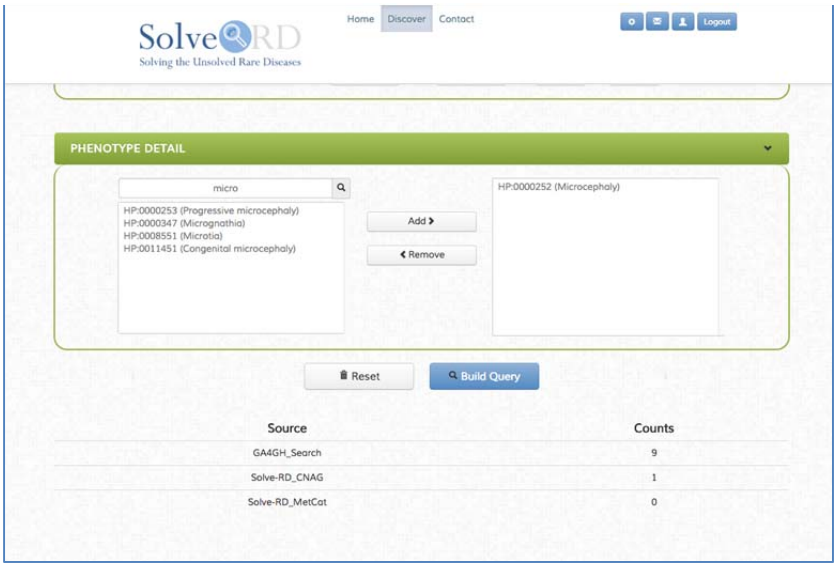

Figure 4. Query Results Display. Results are presented as a count of matching records for each data resource ("source") present in RD-NEXUS. Whether a source and/or counts are visible, or the whether it is possible to click through to data or contact details is dependent on the identity of the user and their assigned permissions.

Conclusion:

Next steps

The RD-NEXUS system is now ready to be presented to ERNs and other Solve-RD partners, along with this Deliverable, so we can explain the concept of federated and bespoke asset discovery (vs asset sharing) and how this might then also be used for advanced forms of patient matchmaking. We will thereby seek their views about if, how and when they would like to employ this approach within research and clinical endeavours. This consultation will be done in conjunction with the ERN Research Working Group, which was established for precisely this kind of purpose (phone and email discussions have taken place, and first face-to-face meeting is planned for November 2018). We anticipate that several ERN partners that have already deposited data within GPAP may then give permission for us to represent obfuscated versions of (some of) those patient records within the RD-NEXUS. In parallel, we will replicate the whole platform on a server that has been made available at the Academisch Ziekenhuis Groningen (UMCG) partner, in order to establish a 2 member federated RD-NEXUS network (using capabilities already built into the underlying Cafe Variome technology). Other sites will be added to this network as interest in the system grows.

As a mixed community of users, data owners, and technical developers, we will then be in a strong position to undertake use-case driven improvements to the whole system. Specifically, we will work on topics such as:

- Creating simple ways for ERN members to enter safe levels of case data into the discovery / matchmaking ecosystem, whilst retaining full control over its discovery and with no data sharing taking place
- Adapting query options, data obfuscation processes, and user groups according to the needs of different use cases
- Enabling cross-communication with other discovery systems and APIs, with an initial target for this being the RD-Connect biosample catalogue

- Exploring more advanced federation and network architectures, not least to allow more flexibility over data models via dynamic adaptation to having different data fields at different locations
- Working with GA4GH to define a rich set of Findable Facets for their Search API, and eventually switching RD-NEXUS over to that once it is powerful enough
- Installing private RD-NEXUS networks with and for a number of ERN consortia
- Publication and dissemination of our progress
- Feeding this work into the EJPRD, as part of that project's wider ambition to create a holistic RD discovery ecosystem as part of the FAIR mission

Appendix 1: Overall RD3 Data Model v1.0