

4 Knowledge Management

4.1 Summary

IMI aims to boost Europe's biomedical R&D base by improving the predictability of safety and efficacy evaluation during the drug development process. The Knowledge Management (KM) pillar is an essential component of IMI that provides the data-pooling and data processing infrastructure to support IMI public-private collaborations in Europe.

The IMI Safety and Efficacy pillars, through a process of calls for proposals and project evaluation, selection and funding, will initiate and manage a number of biomedical research projects and communities of experts. The IMI Knowledge Management pillar supports the Safety and Efficacy projects and communities of experts with their information management and information sharing, modelling and simulation tasks.

The recommendations concerning IMI Knowledge Management are:

Set-up a **Translational KM** team to co-ordinate and provide the support to the Safety and Efficacy projects, both during the preparation of calls and during project execution.

KM Translational support to the Safety and Efficacy projects can be summarised as follows:

- Expertise in bridging the safety, efficacy and knowledge management scientific worlds, expertise in data and systems integration technologies, awareness of international and EU projects in the IMI related areas, be proactive in engaging with the IMI project organisation, capable of scoping a domain of knowledge and laying out the current state of the art;
- Effective interfacing of knowledge and skills in a number of domains that recur in the Safety-pharmacovigilance and Efficacy pillars, such as molecular imaging, tissue/bio banking, Health Information Technology (HIT) and Electronic Health Records (EHR), bioinformatics, biomarker databases and systems biology.

Set-up a **KM Platform** team that conceives the overall architecture and delivers an integrated biomedical data platform and interactive scientific exploration tools.

The KM Platform is an integrative tool that assures synergies with management and exploitation of research results by bringing data together in an open and consistent format that is suitable for overall data analysis. The creation of such a platform can lead to new biopharmaceutical insight through extensive data sharing.

The KM Platform team will publish project calls that address the development of components of the KM Platform that are currently lacking or that need specific biomedical extensions. The evolving KM Platform will, over time, trigger and support new types of joint project that exploit the availability of new IMI data.

The scientific and functional requirements for the KM Platform can be summarised as follows:

- Data federation: seamless search and navigation across heterogeneous data sources, both private and public;
- Data integration: the capacity to pool data from heterogeneous sources in a scientifically, semantically and mathematically consistent manner for further computation;
- Shared services: the development, sharing and integration of relevant and powerful data exploitation tools such as modelling and simulation.

The requirements can be met using a distributed/federated, multi-layer, service oriented, and ontology-driven architecture. However, severe gaps were identified in the area of data representation and exchange standards, ontology development, data protection, and text mining. A set of generic R&D projects is proposed in order to bridge these gaps and meet the requirements:

- Set-up a Specific Content Action to evaluate and propose approaches for building the core KM Platform backbone information architecture / ontology that copes with a.o. pharmaceutical assay data;
- Joint public-private collaboration-led development of IMI KM Platform missing components. Examples are:
 - Enhanced standards for data protection in a scientific web services environment;
 - Models and modules for exposing web services (semantics), scientific services and the properties of data sources so that they can be used or transformed in a semantically, mathematically, and scientifically consistent manner;
 - Knowledge-representation models and data exchange standards for complex systems;

- Domain-specific ontologies for the more detailed level of scientific data and types of relationships between data elements needed by IMI;
- Extensions to the best text mining tools, such as Biomint and e-Biosci, for capturing implicit information about complex processes, as described in patents and the literature;
- Innovative and powerful data exploitation tools, for example multi-scale modelling and simulation, considering and integrating from the molecular to the systems biology level and from the organ to the living organism level;
- Build the core KM Platform database of validated experimental data extracted from the literature and from Safety and Efficacy projects;
- An expert tool (ontology/schema/rules negotiator, services/data negotiator) to guide users through the complexities of the data, data models, simulation and modelling tools and so on in a federated environment.

In addition, from an organisation point of view, the KM Pillar should include:

- An advisory Science Panel that supports the KM teams in applied information technology matters, the ongoing evaluation of the state of the art, and the identification of complementary and synergistic technology R&D proposals. The members have a proven track record in understanding the needs of modern biopharmaceuticals, and in applying information technology to these emerging areas of science;
- One or more task forces to investigate and report on cross-disciplinary aspects (for example modelling and simulation of physio-pathological processes), validate specifications, and align priorities;
- A cross-disciplinary task force to review issues (legal, regulatory, security and data protection (see 4.2.1), ethical, intellectual property) related to data sharing, and propose guidelines and specifications for implementation.

4.2 Introduction

4.2.1 Understanding the Challenge

Figure 27 positions the IMI KM Pillar in the wider context of the interplay between biology, pharmacology, and medicine. The advent of molecular biology as the major driver for medicinal therapy innovation has brought about the fusion of disciplines, resulting in high levels of complexity. At the interfaces, new disciplines arise, such as:

- **Genomics medicine** using personalised medicine, pharmacogenomics, the integration of genomic testing data into electronic health records and into diagnostic and therapeutic decision support tools, for example. These topics are discussed at the EU level by the BioMedical Informatics initiative from the European Commission, Directorate General Information Society;
- **Translational medicine** using biobanks, toxicoepidemiology, secondary uses of anonymised patient data pools, integrating clinical trial systems with electronic health records (EHR), genomics enabled pharmacovigilance, and so on;
- **Biopharmaceutical R&D** using target and biomarker identification and elucidation, toxicogenomics, predictive safety and efficacy, systems biology, experimental medicine, traditional assay and omics assay data integration supporting integrated modelling, simulation and drug candidate decision making, and so on.

The IMI KM teams support IMI work in all three areas defined above. Moreover, a core set of common terminology, ontology, data interoperability arrangements and the like is gradually building up as stakeholders are increasingly willing to align with defined interoperability requirements. The Knowledge Management teams will support the Safety and Efficacy teams with their data interoperability needs. As a matter of policy, the KM teams will identify those initiatives and standards that shape the common core, and will recommend their use and point out collaboration opportunities.

In this model of the fusion of disciplines, it appears that data cross borders frequently. Effective security and data protection methods and practices are, however, a condition *sine qua non* for that to happen. When patient data are involved, as is the case in areas such as pharmacovigilance, data pooling from electronic health records and biobanks, the strictest safeguards must apply. The data handling and data sharing in IMI projects will be subject to mandated and audited policies and guidelines derived from best practice evaluation. How to implement these in advanced technologies may have to be the subject of further technical research and development.

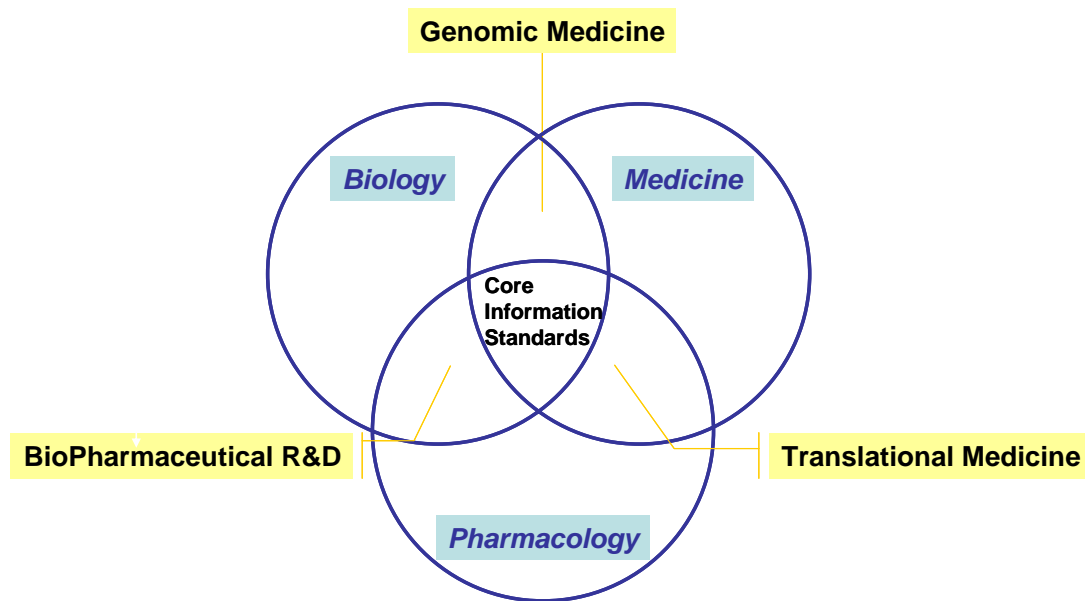


Figure 27 : Positioning the IMI Knowledge Management

4.2.2 Addressing the Challenge

The expert team responsible for creating this section of the IMI Strategic Research Agenda performed an analysis of the IMI scientific and functional requirements, with a series of workshops held in Brussels and Oxford. They discussed current and emerging technologies capable of supporting those requirements, including infrastructure, data resources, data representation and exchange standards, and ontologies. They discussed the nature of the KM services that would be most supportive to the scientific and functional needs of the IMI Safety and Efficacy pillars.

The proposal is to establish a Knowledge Management Implementation body within the IMI Executive Office, composed of two teams. One, Translational KM, will focus on supporting the Safety and Efficacy research projects. The other, KM Platform, will focus on delivering the over-arching integration platform.

The **Translational KM team** will co-ordinate and provide the support to Safety and Efficacy projects during both the preparation of calls and project execution. It will define standards of compatibility across the Safety and Efficacy projects, will promote the sharing of suitable KM technology, and will provide the bi-directional context for Knowledge Management technology R&D, in other words providing a bridge function to the KM Platform team defined below.

The **KM Platform team** will conceive the overall architecture for an integrating biomedical sciences platform, and will research and develop the missing pieces of functionality necessary for data pooling and scientific modelling. It prototypes and tests the evolving platform functionality, and provides the necessary infrastructure. The KM Platform will bring together multi-scale biomedical data on demand for specific biomedical data mining, educational, and modelling purposes relevant to IMI.

This flexible organisation of IMI Knowledge Management activities assures both close support for the Safety & Efficacy scientific projects and the well-co-ordinated implementation of the KM data sharing and modelling and simulation strategy.

At the IMI Executive Office level, the management team includes the leaders of the other three IMI pillars. That Executive Office management team will provide high-level co-ordination, attention for cross-disciplinary aspects, cross-validate project call level documentation, and align and synchronise activities to focus on priorities. These Executive Office management team interactions assure that the IMI IT infrastructure is properly dimensioned, and its functionality closely aligned with the business needs, the scientific requirements and with the priorities of the Innovative Medicines Initiative.

On demand, the KM Science Panel advises the Knowledge Management teams on matters related to the technical architecture, the technical merits and clarity of content of the calls for proposals, the most productive time sequence for the development of the various KM Platform components, and the Knowledge Management technologies in general and their applicability to the biopharma and biomedical sciences.

The responsibilities of the Knowledge Management Implementation body within the Executive Office are to implement the KM SRA.

The Translational KM team will be responsible for:

- Ensuring that the Safety and Efficacy calls for proposals include the opportunities for collaboration with existing KM projects and the information on KM standards;
- Participating in the organisation and evaluation of the Four Pillar project proposals through a quality peer review process, and support the process of arriving at recommendations for project funding;
- Establishing and maintaining a network of expertise in the application of informatics to the needs of predictive safety and efficacy. Building and maintaining an inventory of expertise, methodologies, education and training;
- Providing leadership to Education and Training in the area of biomedical knowledge management talent development;
- Identifying opportunities and managing the exploitation of the databases and the KM Platform.

The KM Platform team will be responsible for:

- Developing and submitting the KM Platform calls for proposals through a consultative process;
- Managing the KM Platform evaluation of proposals and granting of projects;
- Conducting an extensive study of the international and European state of the art prior to the KM Platform calls;
- Managing the KM Platform projects;
- Supervising the organisation and availability of the IMI scientific IT infrastructure.

Figure 28 below illustrates how the Translational KM team will provide this support and play its bridging role.

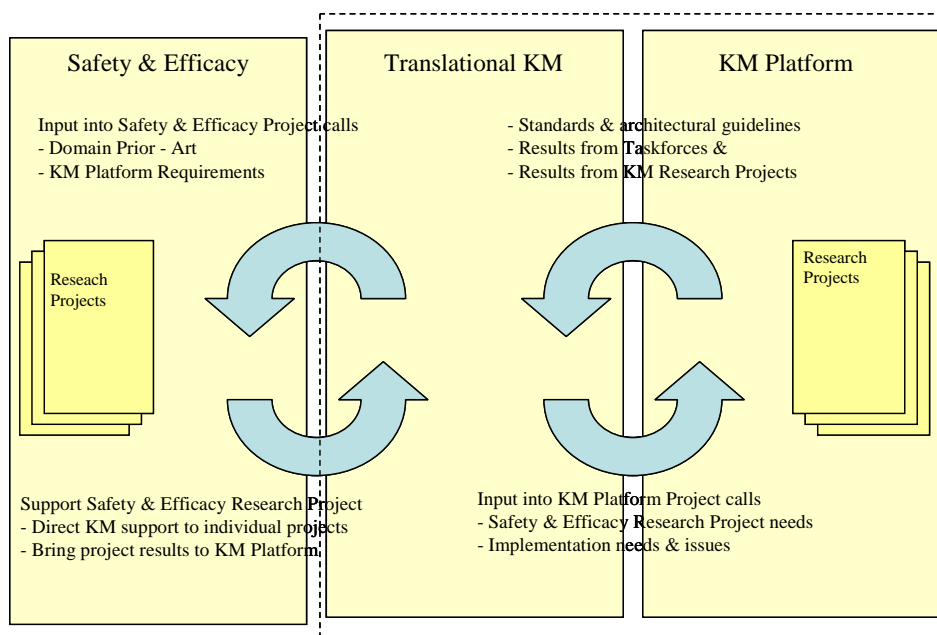


Figure 28 : Knowledge Management Embedding and Bridging

4.3 Translational KM

The Safety and Efficacy research projects will have an information management and data modelling component that is integral to the project and its deliverables. The Translational KM team will support these projects at two different points. First, when the call for proposals is written by documenting the state of the art, by identifying opportunities for leveraging outcomes from existing EU initiatives and specifying the required standards and conventions. Second, during the execution of the project, to ensure project success in implementing the standards and the integration of project results into the Knowledge Management Platform.

It is equally important that specific projects (Safety and Efficacy), with their own IT support, are co-ordinated with the overall KM Platform strategy of IMI and with the evolving capabilities of the KM Platform. It is the task of the Translational KM team to ensure this co-ordination happens.

Each Safety and Efficacy project with an informatics component will fully fund and resource its information management needs. This means that, with respect to the Knowledge Management endeavour, the projects will:

- Fund their IT infrastructure, preferably making use of the KM infrastructure arrangements;
- Fund the IT components, such as gateways, shadow servers, SOA servers and so on, that are required to integrate the project results into the KM Platform;
- Fund the human resources dedicated to the KM team that are needed for the KM liaison function and for integrating their project outcomes into the KM Platform. The number of people involved will vary depending on the complexity of the tasks, as described in the project's programme of work. At a minimum, this should be one FTE.

Examples for Translational KM joint projects are described below. The three examples that follow are taken from this SRA.

4.3.1 Biobanks

The IMI Safety and Efficacy pillars request the availability and integration of biobanks:

- In chapter 2, Improved Predictivity of Drug Safety Evaluation, the organisation of human tissue banks is recommended to address research into intractable toxicities;
- In chapter 3, Improved Predictivity of Efficacy Evaluation, human tissue banks linked to medical records data on phenotypes is considered as a 'must-have' capability. Such banks would serve EU disease-specific regional biomarker centres for the validation of omics based biomarkers;
- In chapter 3, the section on Brain Disorders requests tissue banks for their biomarker work. The section on Inflammatory Diseases requests co-ordinated networks of tissue banks. And in the Diabetes section, biosample banks are requested for biomarkers identification and validation;
- In chapter 3, the section on Cancer, establishing a Systems Biology platform relies on the analysis of tissue samples, body fluids and their data from biobanks;
- In chapter 3, biobanks are also required to couple samples data to medical or clinical data in order to link molecular targets for drug intervention with the pathophysiology of disease, and translate the results of clinical trials into the molecular understanding of the responses;
- In chapter 3, under patient recruitment, the quality of patient data is supported by first-class patient records and biobanks allowing intelligent patient selection, and allowing the investigation of the basis of response and non-response.

A Biobank is a longitudinal and large collection of well-defined and curated biological samples plus the related data including clinical and molecular data. Biobanks feed back into basic research relevant biological samples and their associated pathology data and molecular characterisation data, plus treatment and clinical outcomes data. They are part of the bridge between research and clinical practice, and are an element of the emerging Translational Medicine discipline. They are also addressed in the US NIH Roadmap⁴⁴.

The demand for samples (for example DNA, RNA, proteins, cells, tissue, blood and fluids) is thus increasing for use in high throughput genomics analysis techniques such as sequencing and genotyping, expression analysis. Safety and efficacy sciences require large, centralised repositories to be developed wherein the samples and the data are collected from a network of donating sites, including academic medical centres and community hospitals. An example is the US collaboration to conquer cancer 'C-Change blueprint for a National Biospecimen Network'⁴⁵. The public and private sector must, however, engage in an unprecedented level of collaboration to meet this demand. Bringing together and maintaining diverse samples and their related multiscale data is a real challenge, and will require the use of KM tools for the integration of clinical and biomedical data.

The Translational KM team will support the Safety and Efficacy pillars with timely in-depth state-of-the-art input at the time of writing the calls by identifying complementary or synergistic areas of development. It will provide continuous, effective support to the individual research projects, and also co-ordinate the integration of project data into the overarching KM Platform as required.

Elaborating on the KM support for individual Safety and Efficacy Public-Private Collaborations, one would expect that these projects have expressed, planned and financially resourced needs that the KM team

⁴⁴ <http://nihroadmap.nih.gov>

⁴⁵ <http://www.cchangetogether.org>

can then address with pragmatic best practices to provide ways of setting up or integrating biobanks. This may include the use of existing standards following a minimum dataset approach, biobanks publishing their own ontology and data maps to enable spot integration, and adding unique identifiers to the network, service-oriented architectures and so on.

4.3.2 Healthcare Information Technology and Electronic Health Records

The IMI Strategic Research Agenda puts significant value on the availability of quality electronic health records (EHR) across the EU for the purpose of improved pharmacovigilance and improved clinical trial design.

- In chapter 2, Improved Predictivity of Drug Safety Evaluation, in the Pharmacovigilance section, the creation of an improved clinical trial capacity and capability across the EU is required. In this context, data sources must be enriched with epidemiological, exposure and outcomes data for trials, clinical practice, home care and hospital care. It suggests that there should be technology standardisation in electronic patient records, thus supporting data pooling and integration;
- In chapter 3, Improved Predictivity of Efficacy Evaluation, a pan-EU infrastructure for clinical trials is required to establish an intelligent environment that supports the creation of the electronic patient database of the future. This environment will allow clinical and experimental data to be correlated for the study of biomarkers;
- In chapter 3, to improve patient recruitment, a first-class electronic patient records is required;
- In chapter 3, to improve communication with regulatory authorities, the programme foresees the collection and pooling of benefit–risk data about medicines from patients' medical records;
- In chapter 3, in the Inflammatory Diseases section, a pan-EU database of patients with defined uniform diagnostics, including patient history data, is requested in order to support both research and the development of national patient networks and databases.

The healthcare stakeholders are increasingly willing to align their policies and approaches to health information technology (HIT). In the US, this is exemplified by the Health and Human Services (HHS) Office of the National Co-ordinator for Health Information Technology (ONCHIT). In the EU, eHealth activities are mainly organised through the Directorates General of Information Society, Research, and Health and Consumer Protection. The EU's eHealth Working Group and eHealth Stakeholders Group, which were established in 2005, oversee the implementation of EU eHealth Action Plans. These plans are the result of the i2010 European Information Society programme, and its specific health-related 2005–08 workplan. In addition, the European Commission's FP6 programme supports eHealth and bioinformatics R&D and implementation through the RIDE and Artemis projects.

The eHealth provides a platform for IMI to involve and collaborate with health information technology providers.

An electronic health record (EHR) is a complete set of data across a lifetime about a person's past, current, and prospective health status. It also includes the healthcare that has been provided or is planned. It is stored in a coded, structured, machine-readable form, in multiple jurisdictions and locations, and is accessible in its entirety or in part to legitimate users such as providers, allied health services, emergency services, patients and researchers, from one access point, anywhere and anytime. Patient physiological, clinical laboratory, genomic, environmental factors and treatment and prescription data come together in this virtual file. Provided the necessary safeguards for patient protection and due process (see 4.2.1) are in place, it is an indispensable data source for translational medicine. However, the way that genomic data will be captured in the electronic health record is still an area of research, and one that will be of much interest to the IMI endeavour. How the secondary use of health information data for research purposes will be organised in the future will require a great deal of involvement and attention for privacy and ethical reasons.

Integrating clinical trials, in other words trial protocol driven care, in EHR systems is an area of systems development where the biopharmaceutical industry can bring unique perspectives, needs and expertise to the table, and can take a leadership role. Standards will play a crucial role here too. The CDISC (Clinical Data Interchange Standards Consortium) is a global, vendor-neutral, platform-independent data standard for information systems interoperability. It supports the acquisition, exchange, submission, and archive of electronic clinical data. Participating in its development are global biopharmaceutical companies, technology and service providers, academia, regulatory agencies, and others.

The Translational KM team will assist the Safety and Efficacy pillars with extensive professional networking within the broader EU eHealth community and with state-of-the-art work where the collaboration opportunities have been clearly identified and reflected in the call for proposal documents. The Translational KM team will, equally, actively support the individual Safety and Efficacy Public–Private Collaborations

that have expressed, planned and financially resourced needs with their health information technology integration work.

4.3.3 Biomarker Databases and Data Integration and Analysis

The IMI Strategic Research Agenda requests that an integrated data package standard should be developed to support the acceptance of biomarkers by regulatory authorities in drug filings for new therapies. This requires leveraging data from disparate omics technologies to enable the data to be analysed and mined in integrated and predictive ways:

- In chapter 2, Improved Predictability of Drug Safety Evaluation, this means the creation of a Bio-pharma data warehouse, with data shared by the pharmaceutical companies about GLP toxicity studies on both new and terminated compounds. The same data warehouse will store the data from the IMI predictive safety research and from any other data sources to support the development of *in silico* models of toxicology that are widely applicable;
- In chapter 2, pharmacovigilance depends on improved data resources – epidemiology, exposure data, outcomes data, medical product data, genomics data and so on – harmonised by a pharmacovigilance ontology and analysed with novel signal detection and data mining tools in order to improve the evidence base and benefit–risk assessment methods;
- In chapter 3, Improved Predictability of Efficacy Evaluation, disease specific biomarker registries and pharmacomedicinal databases are requested to support the biomarker validation work;
- In chapter 4, Knowledge Management recommends building a KM Platform that integrates the validated results from the above-mentioned Safety and Efficacy research activities, and provides the tools to search for and analyse occurrences of a specific scientific interest. The breadth and depth of the KM Platform support the elucidation of additional areas of important research, including the systems biology endeavours.

Systems biology is the quantitative study of biological systems that enables computational analysis of the observations to be carried out⁴⁶. Its goal is the predictive understanding of the whole biology. Systems biology is needed in order to understand biological systems at the predictive level, as required for disease detection, prevention or cure.

The KM Platform described in the next chapter will further the state of the art in this area of large-scale data integration and multiscale modelling and simulation. Many individual safety and efficacy projects will, however, endeavour to develop smaller-scale data compendia and analytical tools that support a specific research need in predictive toxicology or predictive efficacy. These will be supported by the Translational KM team. Examples are the FP6 integrated project in toxicogenomics, CEBS; the FP6 integrated project InnoMed PredTox and the US project Oncomine (cancer transcriptome); and the work being carried out in the EU-funded integrated project BioSim.

4.4 The KM Platform

4.4.1 Introduction

The goal of this chapter is to provide input on the enabling technology – the set of technologies and processes required to process data and information – thus allowing knowledge creation, sharing and reuse. This is required to establish a KM Platform capable of supporting the data- and tool-sharing objectives of the Strategic Research Agenda.

The required flexibility can only be met by a federated, multilayer architecture in which independent components, data sources, scientific services and the like can be configured dynamically and articulated by rules and ontologies. In such a configuration, three areas have been identified as critical:

- Technical infrastructure architecture and services (4.4.4);
- Data sources and properties (4.4.5);
- Knowledge representations and models (4.4.6).

⁴⁶ Zoltan Szallasi, Systems Modeling in Cellular Biology, The MIT Press, 2006

4.4.2 Scientific Objectives

Advanced technologies, such as high-throughput screening, genomics, proteomics and metabonomics, have resulted in data generation on a previously unknown scale. Information derived from these data is extensively used in R&D. Examples include target identification and validation, formulation of hypotheses, identification of specific pathways associated with disease states, diagnosis and monitoring. Data integration across heterogeneous data sources and data aggregation across different aspects of the biomedical spectrum, therefore, are at the centre of current biopharmaceutical R&D.

The KM Platform will, ideally, allow:

- Data to be searched, queried, extracted, integrated and shared in a scientifically and semantically consistent manner across heterogeneous sources, both public and proprietary, ranging from chemical structures and omics to clinical trials data;
- Scientific tools such as modelling and simulation to be integrated and shared as modules in a generic framework, and applied to relevant dynamic datasets.

A tool such as the KM Platform can only be successfully conceived, specified, developed and used meaningfully in close collaboration with the Safety and Efficacy pillar projects. The KM pillar organisational approach which will ensure such alignment and collaboration was presented in section 4.2.2 of this document.

The KM Platform's contribution to the previously described examples will be:

Biobanks: The KM Platform could further the state of the art in harmonisation and interoperability by addressing the issues of inconsistent semantics among biobanks within and across organisations. Co-operation between biobanks is very difficult, and there is a major impact on querying for samples – identifying the correct cases and samples across multiple biobanks is not simple. Major data interoperability efforts are thus needed. Examples include NCI CaBIG in oncology, OESO biobank standards, several EU INFOS and RDT projects, and CONTICANET.

Healthcare Information Technology: The KM Platform team can be directed into joint projects, producing integrated patient data from public and private sources, using the IMI KM Platform.

Biomarkers and Data Integration: There is also a huge reservoir of proprietary data, held by both companies and regulatory bodies, on active and discontinued products, as well as marketed products, covering the full scope of R&D. This includes any data from chemical structures to toxicity studies and clinical trial data. These datasets provide invaluable research tools. They could be pooled, possibly supplemented by data extracted from patents and the literature, to increase the predictive power of current models, to revisit and to improve current models, and to populate newly developed models.

The emerging systems biology approach, aimed at understanding complex physiological and pathophysiological processes, requires both data integration at the molecular level, for example through omics technologies, and the availability of sophisticated mathematical or computational models at the pathway, cellular, organ or disease physiology levels, the so-called multiscale models. Although such modelling efforts are still in their infancy, they are rapidly maturing, and some integrated computational models are already in use.

Relevant models are at the centre of all these scientific endeavours. However, the development of multiscale models is a complex task, which is limited by a lack of integrated scientific knowledge. As a result, developing these models can only be undertaken by joint efforts, by the collaboration of scientists from different disciplines, and by goal-oriented and focused initiatives and projects.

The KM Platform addresses this more general need for data integration and aggregation across many heterogeneous data compendia, covering the whole spectrum of biopharma and pharmacomedicinal data, and the need for interlinked modelling and simulation tools that can operate on and co-operate on these datasets. Examples are the Canadian biomolecular interaction network database (BIND), US Pharma GKB database, the EU FP6-funded database BioGRID (microarray data and protein interactions), the EU imaging database MammoGRID (mammography images) and the EU FP6 project GEMSS (grid-enabled Medical Simulation Services).

The terminology 'KM Platform' is used to emphasise the purpose of deriving new knowledge from these data integration and aggregation efforts. The terminology GRID has been used in other publications, emphasising the network computing aspects of such platforms. The meaning of the term GRID has evolved from on-demand high-performance networked computing capacity arrangements such as, for instance, EGEE (Enabling GRID for e-Sciences in Europe), to include the middleware for data integration, data aggregation (Data GRIDs) and the tools for knowledge discovery (Knowledge GRIDs). Examples are the disease specific NCI CaBIG (Cancer Biomedical Informatics GRID) and the proposed EU HealthGRID, BioGrid.

Whatever the terminology, one must address the need for a network computing infrastructure based on internet technology standards for middleware-based semantic interoperability, for multi-scale data representation, and for linking investigative modelling and simulation tools in a scientific workflow: the KM Platform.

Furthermore, it will be important to consider collaborations, to peer review the content model, and to address long-term maintenance by, potentially, involving commercial organisations or public institutions that have experience with the publishing of database resources. This could also be done by setting up a virtual organisation that distributes the work and is committed to the long term.

4.4.3 Technical Objectives

The key technical requirements are:

- Flexibility; in other words modularity (supporting integration of new resources in a standardised way) and configurability (accommodating existing and emerging needs). This is required because:
- The *a priori* scientific and functional requirements are broad and diverse;
- The data resources to be federated by the KM Platform are characterised by a deep heterogeneity in terms of source, ownership, availability, scientific content, quality, level of curation, database design, data organisation, semantics and so on;
- The diverse usage for simulation, modelling and navigation using a variety of methods, some of which are likely to emerge as the result of new R&D;
- The complexity of the underlying science, as well as the complexity of applicable knowledge representation schemas and applicable scientific algorithms;
- Intuitive access to information. From the user's point of view, the knowledge management platform must provide relevant and simple access to information – both in terms of searching and navigation – and to services. It must also provide precise organisation of the content, independent of its source, allow scientifically relevant data integration (data pooling) and data exchange, and provide mechanisms for data capture and annotation. In addition, it must provide a dynamically evolving set of validated data exploration, analysis, simulation, and modelling services. Finally, it must be consistent with the way community participants work, and integrate smoothly into their day-to-day environments;
- A collaborative environment. It should provide collective working, virtual meetings, knowledge sharing, forums, discussions and so on, which are open to the communities of experts;
- A toolbox for analysis, visualisation, modelling and simulation.

From the technical point of view, the KM Platform must, therefore, ensure seamless data integration across a broad range of heterogeneous resources; interoperability of computing services and applications (semantic, scientific, and technical) across organisations and networks; secure and robust mechanisms for data and services management; and a flexible, intuitive, collaborative environment.

In principle, the requirements for the KM Platform can be met by designing a federated environment articulating independent tools, components and resources based on open architectural standards, which is customisable and capable of dynamic reconfiguration.

4.4.4 Solution Component 1: Technical Infrastructure and Services

The KM Platform's requirements are best addressed by a distributed/federated, service-oriented, ontology-driven, layered functional architecture:

- Basic IT infrastructure layer: hardware, operating system software, connectivity network and services such as quality of services, data integrity, firewalls, redundant systems, back-up infrastructure, computer clusters and so on;
- The backbone: services providing basic functionality (data access and security) and interoperability (for example messaging and brokering);
- Data access to heterogeneous resources through a data virtualisation layer (decouple data from their local schema and make data access independent of platform and schema) and a data abstraction layer (provide a common view of all accessible data via a set of ontology / rule-mapping mechanisms);
- Services layer, making application services accessible over the backbone and connecting to data resources;
- Connections layer, providing a secure access point to all authorised users and processes;
- Organisations, describing users and allowing them to share data and services, and collect information.

The most appropriate current technology providing the required flexibility is web services and, in some cases, business-to-business platforms. For handling the scientific tasks in IMI, however, current web service descriptions and annotations must be improved.

Security will have to be addressed at multiple levels:

- Infrastructure;
- Application access;
- Data Protection;
- Access control, which would be policy-governed;
- Privacy-enhancing technology, such as de-identification.

Security and privacy are active areas of research, and technologies are emerging that could be used to ensure the security of the platform. See 4.2.1 on security and data privacy.

The KM Platform will organise necessary and sufficient IT infrastructure services to support the infrastructure that is needed for the Safety and Efficacy projects, as well as to support the KM Platform's own IT requirements.

In selecting the IT services, the KM Platform team will consider results from several network computing projects, or GRIDs, which were initiated by the current FP6 programme, some of which target eHealth or biomedical science and practice communities throughout the EU.

The functional architecture discussed above is summarised in Figure 29 below:

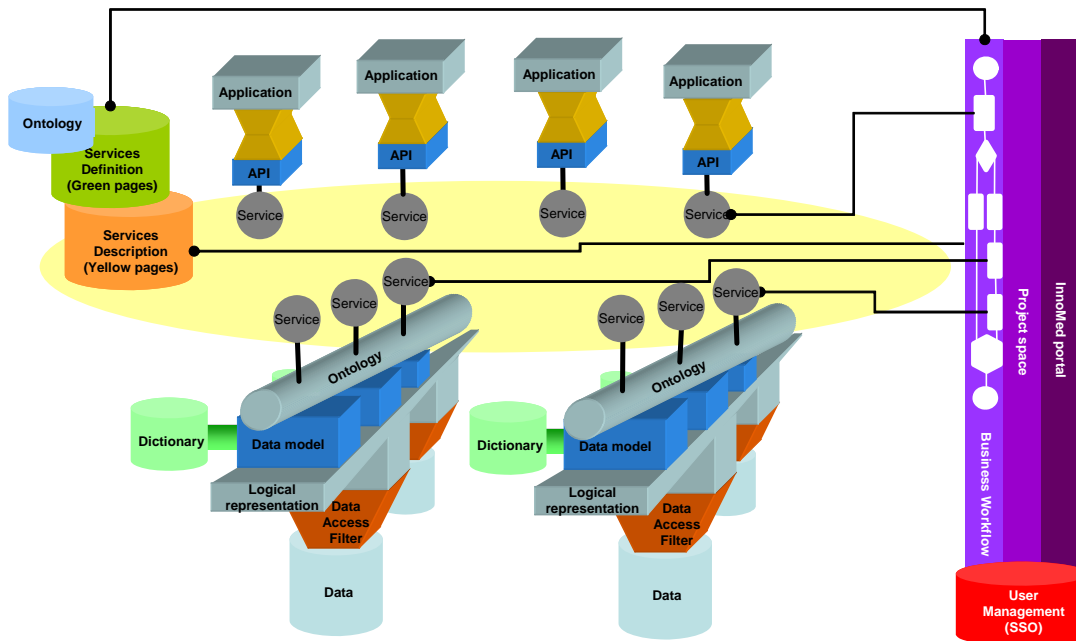


Figure 29 : Knowledge Management Functional Architecture

EU-wide broadband networking is an integral part of the EU's eHealth vision and programme, and the Commission Services initiated the HealthGRID project. The Knowledge Management team will take that concept into consideration when recommending an IMI IT infrastructure and service provider.

The KM Platform team will develop, maintain and publish an IMI IT computing resources infrastructure that provides guidance for consistency, and it will insist with all parties involved that no money is wasted in duplicative efforts or on incompatible ventures in isolated infrastructure islands. There should be only one IMI Data Centre.

4.4.5 Solution Component 2: Data Resources

The data described by the scientific requirements is heterogeneous. It includes:

- Proprietary experimental data, for example from pharmaceutical companies;
- Highly curated, experimental-quality public domain data, such as SwissProt and PubChem;
- Publicly available, qualitative, documentary data such as Medline, WDI and CAS, sequence databases, and chemical structures database, for example CAS and Beilstein.

Provided it is possible to apply relevant transformations to build composite datasets that are consistent in terms of data content, data quality, data descriptions, and of mathematical properties with the scientific objectives and algorithms to be used, it is possible to assemble useful aggregated datasets.

One requisite is that the implicated data sources should be fully understood: how were the data obtained, for what purpose, what are their quality and validation levels, how complete is the dataset, what the dataset's bias, what are the standard errors of the measurements, what protocols were used, and so on. In the data exchange formats that are currently under development, these aspects are poorly developed. There is no mechanism for creating a virtual experimental data warehouse on the fly. Standard specifications should be developed in the areas of omics and clinical trials.

In addition to issues that relate to the meaning and significance of data discussed above, data sources can differ widely in data curation and the quality control that was applied when they were created. This is a particular issue for pre-clinical experimental data repositories. Data quality is critical, however, and sub-standard data must be eliminated. This requires, at the very least, that recommendations should be developed within IMI, and an index of quality (confidence) be assigned.

The whole process of data aggregation should be transparent, and remain under the control of the scientist. The simple 'wizard' approach to guide the user through the possible processes and workflow would probably fail: too many sources, scientific models and algorithms are integrated into the KM Platform. A new type of 'intelligent wizard' will have to be designed.

4.4.6 Solution Component 3: Knowledge Representations

This layer should be based on a set of business entities uniquely defined across IMI data sources and services. This will ensure:

- Reliable and consistent information integration and consolidation across heterogeneous data resources;
- Consistent interaction with the data;
- Interoperability of services, at the semantic and technical level;
- Relevant configuration of the services.

By business entities, we mean an aggregate of data (or a representation) that describes some entity (science object) that exists in reality and is relevant to the Innovative Medicines Initiative scope. Examples include a protein, a tissue sample, an assay, a protocol, a domain actor such as a research unit, a person, or even some information resource such as a document or a technical schema. The collection of business entities is governed by an ontology which describes the business entities and their properties, and the relationships between them. The rules that will ensure quality data collections should be crafted by resorting to these ontologies.

Complex business entities can be assembled from a well defined set of elementary business entities. For example, a business entity describing an assay result will probably comprise a compulsory set of elementary business entities describing science objects such as chemical entity, buffer, dilution, molecular target, protocol, species, strain, unit and so on, each with specific types and properties.

Each business entity is assigned logical property attributes, which define how it can be processed, and descriptive attributes, which define what the business entity is about. Together, these logical and descriptive attributes must be sufficient to describe the data element properties fully and unambiguously, to drive the methods that are applied to the data elements, such as calculations, translations, transcoding and transformations, and to search, navigate, explore, filter, and aggregate data.

These representations and ontologies will be used for a variety of purposes, such as searching and mapping data from heterogeneous sources, dataset navigation and exploration, data aggregation and data

visualisation. They will describe generic relationships, properties, restrictions and constraints independently of any local context. Together, they will form the upper level KM Platform backbone business entity ontology. For the most part, it will have to be developed, taking into account current initiatives, existing standard data representation models, and reference ontologies currently used in the life sciences.

Ontologies used to describe data properties, restrictions and constraints of local data repositories will have to be mapped to the KM Platform backbone business entity ontology. This will require each local data source to expose its local ontologies (and logical schema, rules etc) to the central KM Platform repository via a mapping negotiator, to align and validate the different sources to a consistent composite view of the data (semantically, mathematically, and scientifically), and to configure the connector. The result will be a semantic hub mapping local attributes (plus associated definitions and rules) to the KM Platform core ontology. Similar tools will be required to map the schema of the source database to the data federation tool. The ontology drives an interactive data negotiator articulating data and services. Further developments are needed building on current mediator technology.

4.4.7 Looking for Synergies

Several of the issues addressed above, notably in the area of data integration and semantic interoperability, are the focus of European Communities-funded, large-scale initiatives. These include notably:

- The INFOBIOMED Network of Excellence (NoE), focusing on biomedical Informatics, in particular on the development of methods for clinical and genetic data interoperability and integration and on interfacing tools and technologies used in both medical informatics and bioinformatics;
- Semantic Interoperability and Data Mining in the Biomedicine NoE, also focusing on methods for bridging medical informatics and bioinformatics, data interoperability and data mining;
- More generic projects aimed at the wide-scale adoption of semantic technologies, such as the two Knowledge Web NoE and REVERSE;
- Institutions have been created to deal with ontology in general, both in Europe (Centre for Ontological Research) and the US (National Centre for Ontological Research), and biomedical informatics in particular (IFOMIS);
- Commercial or open-standards organisations active in this field.

Synergies should be identified between the Innovative Medicines Initiative and these organisations, and research efforts should be aligned. Similarly, an inventory of current initiatives on biomedical datasets, representation models, specialized applications, grid computing, semantic grids and the like should be carried out.

4.4.8 Building the KM Platform

The KM Platform is not an end in itself, but a set of tools to advance predictivity in drug safety and efficacy. The KM Platform activities are thus initiated and executed in close collaboration with the Translational KM team and the Safety and Efficacy Public–Private Collaborations. In the preparation phase for the call for proposals, the Translational KM team will contribute with state-of-the-art briefs from which guidance for the prospective Public–Private Collaborations will be derived, and included in the call for proposals documentation. The guidance will cover topics such as current concepts, collaborations to consider, opportunities and standards.

At the most general level, the KM Platform comprises both the infrastructure and multilevel connectivity component, and the toolbox and application component for modelling, simulation and visualisation. The KM Platform advances the state of the art for purposeful multiscale data integration and its scientific exploration, while focusing on predictive safety and predictive efficacy sciences.

4.4.8.1 Specific Content Action

A Specific Content Action will be set up to evaluate and propose approaches for building the core KM Platform backbone information architecture / ontology that can cope with a.o. pharmaceutical assay data. The feasibility and quality of the KM Platform, and ultimately its scientific relevance, relies on high-quality, robust, business-focused, scalable, state-of-the-art ontologies. These ontologies must be built on sound theoretical foundations for the solution to be viable and resilient. We therefore suggest that, as a preliminary to KM Platform development projects, the Specific Content Action should be set up to evaluate all aspects of information architecture, the needs, the standards, the current research efforts and to evaluate what it would take to build the KM Platform backbone and, possibly, to provide proof of concept. The recommendations will guide further KM Platform components development.

4.4.8.2 KM Platform Development Work

The following topics were identified as areas in need of further R&D:

- Review security and privacy issues, notably in the legal, regulatory, ethical and intellectual property areas, and propose guidelines and specifications for implementation in the context of the KM platform;
- The technology for data protection in a Web Services context is not mature. Standards are still evolving, with implementations often falling behind; examples include SAML and XACML. We suggest these standards should be evolved to the level required for IMI purposes, for example semantic rich annotations of web services for service discovery. See 4.2.1 for security and data protection;
- Scientific knowledge representations: IMI KM Platform will rely in large parts on the availability of high-quality knowledge representation models and data exchange standards, which are presently largely lacking, inconsistent, or incomplete for scientific data. The focus should be on developing knowledge representations approaches for complex systems such as systems biology and disease models, as well as research processes;
- Domain ontologies: the KM Platform backbone ontology as well as some new domain-specific ontologies will have to be developed and built on sound theoretical foundations, taking into account current initiatives, existing standard data representation models, and reference ontologies currently used in life sciences;
- Text and data mining: current information extraction techniques are relatively successful at extracting entities and simple pair-wise relationships between entities, for example protein–protein interactions. While this is extremely useful, more advanced tools are needed to extract the implicit information about complex physiological processes required by computational models;
- Data extraction and curation: the quality of data is of key importance. Data curation efforts undertaken together with the Safety and Efficacy Public–Private Collaborations should aim at building a KM core reference database of validated experimental-quality data integrated from various sources;
- Ontology/schema negotiator. In a federated system, each data source is independent and connected to the system via wrappers, used for accessing and retrieving data. An expert tool for exposing the properties (including scientific properties) of local data sources and mapping them to the KM Platform core is required;
- Data/services negotiator: a black-box approach should be avoided and the scientist must remain in full control of the process at all times. At the same time, the interface to the system must be relevant, intuitive and simple to use. This will require the design of a new family of wizards, guiding the user into the complexities of the KM Platform data, data models, simulation and modelling tools in a goal-oriented, scientifically relevant and intuitive manner. This includes the development or enhancement of semantic query languages;
- System biology toolbox and framework for assembling and launching composite analytical, simulation and visualisation strategies.

4.5 Resources

The costs realistically estimated for implementing the recommendations described in this chapter are: €14.9 mn per year for a period equal to the initiative's initial duration (2007–13).

The elements considered in calculating the costs for the KM programme are related:

- To establish the Translational KM and KM Platform teams within the Executive Office;
- To establish the Translational KM team with subject matter experts in applying informatics to critical biopharma & biomedical needs areas. The examples given above were biobanks, electronic health records and biomarker databases;
- To establish the KM Platform team with experts in advanced informatics concepts and implementation;
- To establish and operate the technical advisory KM Science Panel.

In addition, resources have been allocated for the IT infrastructure and support in the Safety pillar (€15 mn) and the Efficacy pillar (€21 mn). It is indeed anticipated that many individual research projects ultimately funded by this initiative will have unique IT/KM needs that will be funded as a part of these budgets. These budgets, therefore, support:

- Safety & Efficacy specific applications development;

- IT hosting by an external vendor, or preferably, through contract with the KM Platform;
- IT costs related to data management within the KM Platform.

In total, €51 mn per year have therefore been allocated to the development and implementation of the Knowledge Management part of this SRA. These bottom-up estimates are of an expected magnitude as the rule of thumb is that 10% of the overall budget must be reserved for information / knowledge management tasks. Further successive refinement of this budget will be undertaken when the priorities are defined, the call for proposals is being written, and the consortium projects reviewed and approved.

In the tables below, costs are shown for the Translational KM participation in joint projects, and for the KM Platform R&D projects. An average yearly cost is calculated from these numbers; however, half of the total costs will be incurred during the first 30 months of the initiative because of immediate infrastructure needs.

For the KM infrastructure costs, the tables below borrow numbers for similar efforts from one of the largest IT Outsourcing Services Provider companies.

4.5.1 Running the KM Teams (Translational KM and KM Platform)

| Project support and management | Costs per year (k €) |
|---|----------------------|
| <ul style="list-style-type: none"> • Eight full time staff + admin loaded costs One Head and seven subject matter experts / project managers | 1,500 |
| <ul style="list-style-type: none"> • Science Panel costs – four meetings, including facilities, and ad hoc consultations | 200 |
| <ul style="list-style-type: none"> • Task Forces (State-of-the-art, best practices, legal / ethical, security and data privacy), three at 130k (five involved, 40 man days, four meetings) | 550 |
| <ul style="list-style-type: none"> • Call for Proposals (expert input, call presentation, peer review, etc.) feedback to Scientific Committee SRA | 300 |
| <ul style="list-style-type: none"> • General support (talent inventory obligations, people networking, contribution to IMI communication, education, prior art studies, conferences, etc.) | 750 |
| Total Project Support and Management | 3,300 |

4.5.2 KM Platform Costs

| KM Platform Infrastructure (hardware and software) | |
|---|--------------|
| Central gateway - One central site and three distant sites – set-up and running costs | 3,700 |
| User software licences | 1,800 |
| Development/analytic platforms (licences, applications & development) | 2,700 |
| Manage and meet the service provider / application service provider | 200 |
| Total Infrastructure and management | 8,400 |

| KM Platform Research Projects | | | |
|------------------------------------|-----------------------------|------------------------------|----------------------|
| Research Project | Duration (months effective) | Total costs (k €) 7 years | Costs per year (k €) |
| Platform 'Specific Content Action' | 18 | 2,000 | - |
| KM Platform development | 48 | 8,000 | - |
| Data curation | 24 | 2,500 | - |
| Demonstrator Validation | 12 | 1,500 | - |
| Build final Platform deliverable | 12 | 1,500 | - |
| Total KM Platform Projects | | 15,500 | 2,210 |

4.5.3 Translational KM Costs

In addition to the budgets allocated to KM in Safety, the KM pillar has reserved € 0.98 mn to support joint projects and € 23.7 mn to support Efficacy KM (infrastructure and support).

4.5.4 Summary of Resources Needed

| Activities | Costs per year (€mn) |
|-----------------------------------|----------------------|
| Running the KM Teams | 3.3 |
| KM Platform Infrastructure | 8.4 |
| KM Platform Research Projects | 2.2 |
| Translational KM Joint Projects | 1.0 |
| Translational KM Efficacy Support | 23.7 |
| TOTAL KM (€mn per year) | 38.6 |

4.6 List of Contributors

| Knowledge Management | | | |
|------------------------------------|-------------------|-----------------|---|
| Stakeholders Group | Last Name | First Name | Institution |
| European Commission | Iakovidis | Ilias | DG Information Society |
| | Norager | Sofie | DG Information Society |
| | Norstedt | Irene | DG Research |
| | Rainer | Bernd-Walter | DG Research |
| Academia | Breton | Vincent | CNRS |
| | Ceusters | Werner | Unversität Saarland |
| | De Moor | Georges | University of Gent |
| | Hofmann | Martin | Frauenhofer Institut |
| | Legre | Yannick | CNRS |
| | Sanz | Ferran | Universidad Pompeu Fabra |
| | Van der Lei | Johan | Erasmus University |
| Companies, Pharma & SME | Barnes | Julie | BioWisdom |
| | Boubekeur | Karima | Roche |
| | Boyer | Scott | AstraZeneca |
| | Claerhout | Brecht | Custodix |
| | Coupet | Pascale | Temis Group |
| | French | Dan | Consider Solutions |
| | Gardner | Steve | BioWisdom |
| | Grandjean (Chair) | Nicolas | Novartis |
| | Lave | Thierry | Roche |
| | Lisacek | Frederique | Genebio/SIB |
| | Mestres | Jean-Christophe | IBM |
| | Michel | André | Aureus Pharma |
| | Peeters | Marc | Roche |
| | Pegler | Geoffrey | Insightful AG |
| Vatant | Bernard | Mondeca | |
| Others | Moquin-Patthey | Carole | ESF – European Medical Research Council |