Fraunhofer Biomedical Research Information, Communication, and Computation Workbench

(BRICC Workbench)

White Paper

Version 1.0

Prof. Dr. Carsten Claussen, Fraunhofer IME-SP
Matthias Dobkowicz, Dipl.-Inf., Fraunhofer IZI-BB
Harry Freitas da Cruz, MBA, Fraunhofer IZI-BB
Prof. Dr. Thomas Illig, HUB/MHH
Dr. Manfred Kohler, Fraunhofer IME-SP
Prof. Dr. Norbert Krug, Fraunhofer ITEM
Dr. Jörg Lehmann, Fraunhofer IZI
Dr. Meike Müller, Fraunhofer ITEM
Dr. Guy Redding, Fraunhofer IZI
Dr. Christina Schröder, Fraunhofer IZI-BB

Corresponding author:

Dr. Christina Schröder
Fraunhofer Institut für Zelltherapie und Immunologie (IZI)
Institutsteil Bioanalytik und Bioprozesse (IZI-BB)
Am Mühlenberg 13
D - 14476 Potsdam-Golm; Germany
Tel. +49(0)331 / 58187 227
E-Mail: christina.schroeder@izi-bb.fraunhofer.de

About Version 1.0

With this White Paper, the authors have presented their thoughts and ideas to the scientific community regarding set-up of the BRICC Workbench. They seek discussion and support, and aim to formulate a joint project proposal within the next quarter of the year.

Potsdam, Hamburg, Hannover, and Leipzig, March 31st, 2015

Contents

Executive Summary

Today, biomedical research requires networked consortia and approaches that cut across the borders of institutions, nations, and the public and private sectors. Hence intelligent solutions for secure, privacy-, and confidentiality-preserving data sharing and analysis are of paramount importance for patients to eventually meet their - so far unmet - medical needs, and for industry to reduce the time-to-market for health innovation.

With the BRICC Workbench, Fraunhofer seeks to build Europe's utmost

Industry-centric biomedical research infrastructure,

enabling secure and enduring **data sharing between academia and industry**, and **between pharmaceutical and biotech companies** for early discovery and pre-competitive research, to critically support translation of

- research in human molecular biology and medicine into better drugs and diagnostics, and
- feeding clinical findings back into better hypotheses for research in molecular medicine.

Data streams shared between participating companies and institutions will be secured under highest applicable international ethical, legal, and quality standards, by latest technology, and inspired by a flexible Cloud-oriented Service Market Place (CSMP) concept.

By continuing close collaboration with European Research Infrastructure Consortia (ERIC; e.g. ECRIN and BBMRI), the Innovative Medicines Initiative (IMI) Joint Undertaking, the International Organization for Standardization (ISO) Technical Committee 276 "Biotechnology", and national public and private research initiatives, the Workbench will complement and support these groups' ongoing work by enabling tailored services and applications that fulfill industrial research requirements. For instance, all services and processes available over the Workbench will be validated and certified to support users' subsequent submissions for marketing authorization.

Drawing upon their previous work and expertise, the Fraunhofer institutes IZI, IME, and ITEM will join forces to lay the foundation of an information infrastructure that caters to the pharmaceutical, biotech, and ICT industries, as well as the biomedical research community as a whole, to foster progress in medicine for the benefit of patients and society on a global scale.

Zusammenfassung

Die moderne biomedizinische Forschung erfordert eine breite Institutionen- und Länderübergreifende Zusammenarbeit akademischer Forschergruppen mit der Industrie. Dazu sind intelligente und sichere Lösungen für Datenaustausch und -verarbeitung erforderlich, die zugleich allen ethischen und datenschutzrechtlichen Vorgaben genügen und die Vertraulichkeit der beteiligten Forscher wahren. Nur so lässt sich die Entwicklung neuer Therapien und Diagnostika entscheidend beschleunigen - zum Nutzen der Patienten, der wissenschaftlichen Gemeinschaft und auch der Industrie.

Mit der BRICC Workbench wird Fraunhofer daher in Europa eine

Industrie-orientierte Infrastruktur für die biomedizinische Forschung

aufbauen. Diese wird den sicheren und dauerhaften Datenaustausch sowohl zwischen Hochschule und Industrie als auch zwischen Pharma- und Biotech-Unternehmen untereinander unterstützen, damit im Sinne der „translationalen" medizinischen Forschung

- aus molekularmedizinischen Forschungsergebnissen schneller neue Therapien und Diagnostika entwickelt sowie
- aus klinischen Daten bessere Hypothesen für die molekularmedizinische Forschung generiert werden.

Über eine flexible Plattform - etwa nach dem „Cloud-oriented Service Market Place" (CSMP) - Konzept - werden die beteiligten Unternehmen und Institute ausschließlich nach international gültigen ethischen, rechtlichen und Qualitäts-Standards auf die relevanten Daten zugreifen können; der Zugriff wird mit neuesten Verfahren gesichert.

Mit der BRICC Workbench ergänzt Fraunhofer seine bewährte und enge Zusammenarbeit mit den "European Research Infrastructure Consortia" (ERIC; z.B. ECRIN und BBMRI), der „Innovative Medicines Initiative (IMI) Joint Undertaking", dem Technischen Komitee 276 "Biotechnology" der Internationalen Standardisierungs-Organisation (ISO) sowie mit zahlreichen Forschergruppen in Industrie und Hochschule um Forschungs-Dienstleistungen, die insbesondere auf die industrielle Forschung zugeschnitten sind. So werden z.B. auf der BRICC Workbench ausschließlich qualitätsgesicherte, validierte bzw. zertifizierte Dienste zur Verfügung gestellt, die in die Zulassungsunterlagen z.B. für Arzneimittel oder Medizinprodukte einfließen können.

Die Fraunhofer-Institute IZI, IME und ITEM werden ihre langjährigen einschlägigen Vorarbeiten und Erfahrungen gemeinsam nutzen, um jetzt den Grundstein für eine IT- und Service-Infrastruktur zu legen, die industriellen Anforderungen genügt. Sie wollen damit nicht nur der wissenschaftlichen Gemeinschaft einen entscheidenden Dienst erweisen, sondern vor allem den medizinischen Fortschritt zum Wohle der Patienten und der Gesellschaft insgesamt befördern.

1. Background

Whilst generating ever-increasing data volumes and complexity, biomedical researchers have promised society for more than a decade that basic research results from human molecular biology and medicine would soon translate into better drugs and diagnostics ("translational", "personalized", "individualized" or "precision" medicine), and that clinical findings would vice versa inform better hypotheses for research in molecular medicine ("from bench to bedside and back"; (1)). To deliver on these promises, a comprehensive infrastructure is required that promotes interoperability, computation and analysis of all the relevant genotypes and phenotypes of patients and healthy individuals, of drug targets and compounds, on a common platform

- across the entire spectrum of diseases,
- across all participating bodies (e.g. universities, companies, hospitals), and
- across national borders.

Although numerous costly national[1] and international public funding programmes (2) for networked research in translational medicine are underway, many of the funded projects lack appropriate QM and QC measures to secure quality of data, management, and specimens. In addition, despite a wealth of guidelines, best practices and specifications (e.g. in the biobanking or in the informatics field), international quality standards are - though now evolving – far from being widely implemented in academic clinical and omics research. Hence many biomarker findings have historically proven to be irreproducible (3) and/or irrelevant for industrial R & D.

Despite several initiatives to support academic biomedical research in public-private partnership (e.g. 2006 CRIP (4), 2008 IMI (5), 2010 m^4 (6)) , pharmaceutical and biotech industry have up until this point failed to engage over a comprehensive, quality-assured, enduring, sustainable, and scalable common data hub and research infrastructure with academia. Both hospitals and pharmaceutical companies are still facing multiple problems in data sharing with external partners, due to lack of data interoperability and semantic integration tools, skills, and resources. The tranSMART platform (7) which is offered and promoted over the „Innovative Medicines Initiative" (IMI) to leverage data sharing between the said partners, remains - though well advanced - still unmanageable for many academic research groups (who cannot afford external assistance to implement ICT tools) and, moreover, needs in our view to be complemented with additional tools and services.

Against this background, and across numerous projects, the three Fraunhofer institutes IME, ITEM, and IZI, have accrued deep knowledge of current infrastructure relevant for pharmaceutical R & D. Thus they have been able to achieve widespread expertise and implement several key solutions in the Big Data Environment that will critically support the set-up of the BRICC Workbench as tomorrow's biomedical research infrastructure:

- Fraunhofer IZI-BB
  Building metabiobanks based on the CRIP concept for a decade (see Fig. 1 for a recent example (8)), IZI-BB has gained broad and longstanding expertise in trans-institutional data integration and harmonization in the clinical and academic context. Their modular up-to-date software portfolio „CRIP toolbox"(9), including its CDISC compliant semantic integration tools, is ready to support data integration across the entire biomedical research field. Underpinned by the CRIP Privacy Regime (10), the group's solutions comply with all relevant ethical and legal standards, including BBMRI's recommendations for federated biobanking infrastructures (8). In 2013, CRIP was shortlisted by the STRATUM consortium (11) as a possible solution for UK biobanking infrastructure.

[1] Deutsche Zentren für Gesundheitsforschung, Nationale Kohorte, Dt. Knoten für BBMRI

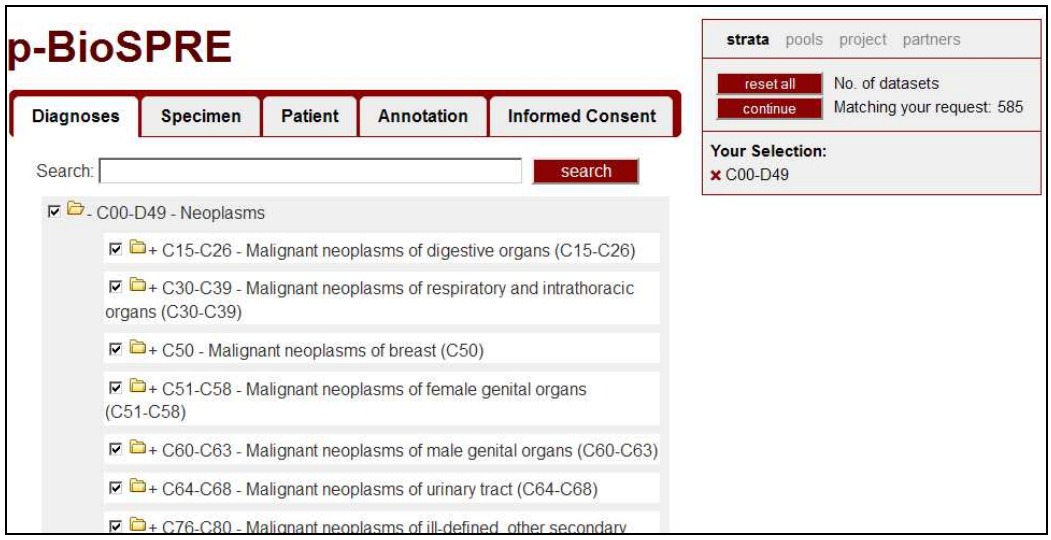

Figure 1: Query Tool of the metabiobank p-BioSPRE
(see https://preview-crip.fraunhofer.de/intern/demo/p-biospre for a demo version)

- Fraunhofer IME-SP

In the IMI program ND4BB, Fraunhofer IME-SP has successfully established an Information Centre together with partners from industry (Fig. 2). It enables legacy and newly created data related to drug development from EFPIA and public partners to be combined, and includes tools and corresponding guidelines, based on state-of-the-art standards for collaborative research and development of new drugs. These tools and guidelines are developed independently of the therapeutic area and can be adapted to additional domains, requirements and/or new standards.

The core translational part of the information repository is tranSMART based (7), and was set up together with two SMEs, The Hyve and GRITsystems. The tranSMART platform is widely used in academic and industrial translational research organizations, in a number of IMI and EU projects (a.o. u-BIOPRED, RA-AMP, OncoTrack, ABIRISK, EMIF and PreDiCT-TB), and it also is the core system of the eTRIKS (European Translational Information and Knowledge Management Services) platform.

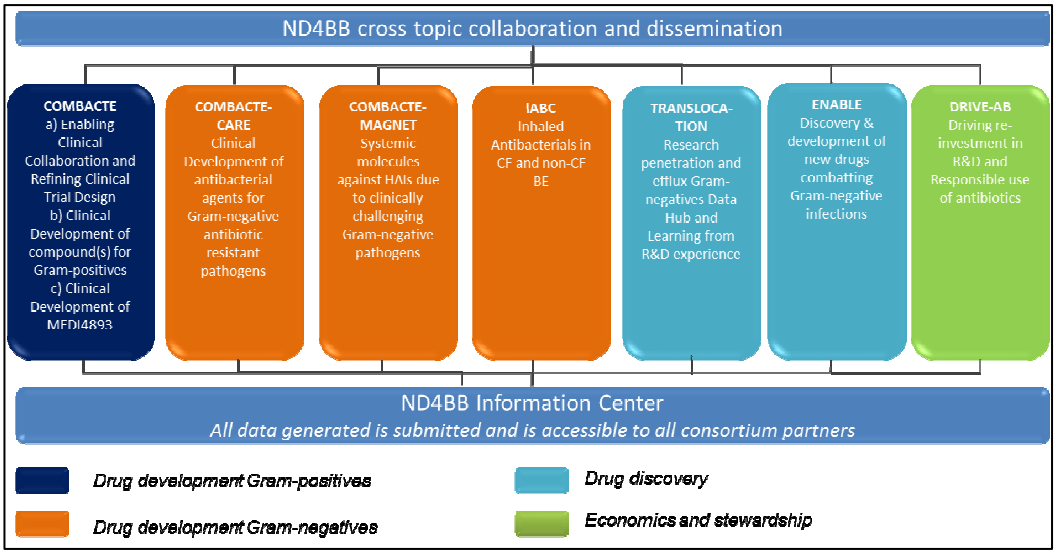

Figure 2: Overall Architecture and Information Centre of the ND4BB Program (as of 2014)

The information repository is combined with a pre-clinical data repository (LSP) and will provide a framework to govern data sharing with highly granular (individual level) security and access/privilege controls (such as read, create, update, delete) which provides hierarchal levels of access to different parts of the systems content. Project participants will get full read access to the information repository along with instructor-lead training and supporting documentation. A Collaborative Access approach (ColA) could give externals escorted access to data. An easy to use web-based system will allow external users to request escorted data analysis procedures on data in the information repository.

- Fraunhofer IZI

Laboratory Information Management Systems (LIMS) are often confronted with demanding requirements involving processing and storage of large amounts of semi-structured data. The Ribolution Data Management System (DMS; (12)), developed at Fraunhofer IZI, is a LIMS that captures the lifecycles of the many thousands of samples and experiments conducted during the Ribolution project using scientific workflows. Scientific workflow governs the experimental procedures on all samples and forms the foundation upon which clinical samples are connected to measurement data. The Ribolution system places safeguards on data input and data processing to guarantee overall data integrity, and eases the arduous task of sample auditing by providing visualisations of sample lifecycles. On top of managing experimental and sample data the Ribolution DMS also captures sample quality and provides quality reports.

Biobanking infrastructure must naturally remain generalised, whilst laboratory infrastructure often becomes specialised. The Ribolution DMS is an example of a specialised LIMS that improves laboratory worker productivity by managing all aspects of sample lifecycles, but it was not implemented with any general requirements involving biobanking infrastructure. Integrating with the BRICC Workbench will provide the means to validate techniques and methods that will be required by users in order to connect the BRICC workbench with services offered by specialist laboratory systems.

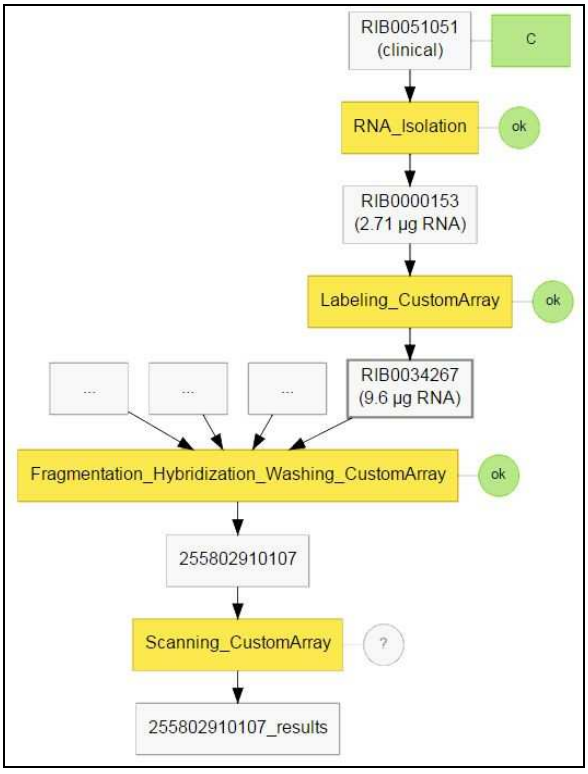

Figure 3: Workflow of Ribolution LIMS;
yellow boxes: lab process steps;
green circles: QC steps

- Fraunhofer ITEM

  Beginning with toxicoloy studies in the 1980s, and encompassing the latest developments in biobank and clinical trial design, ITEM will provide three decades of hands-on experience and expertise with pharmaceutical research for a cutting-edge, yet down-to-earth approach to the BRICC Workbench.

  The Registry of Industrial Toxicological Animal Data (RITA) is a qualified expert system for centralized and standardized storage of results of toxicology studies. For more than 25 years, RITA has supported the evaluation of effects in historical control animals for comparison to actual results. ITEM's expertise in securely sharing data between some 15 pharmaceutical companies from the US, Europe and Japan over RITA is unique: RITA is the only toxicological database worldwide where peer-reviewed data is collected and shared between the participating companies. The long-term experience in handling this complex qualitative data (Fig. 4), combined with established GLP-compliant procedures will certainly benefit set-up of the BRICC Workbench.

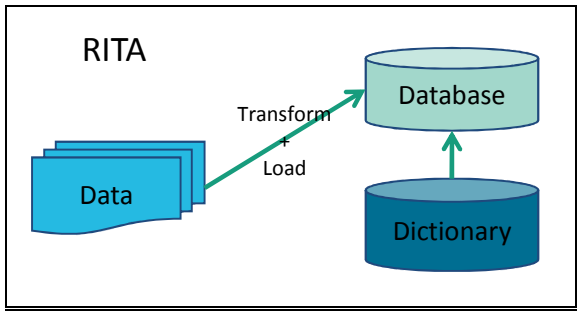

Figure 4: Data transform and load process in 25 year-old RITA system: Direct connection of Dictionary and Database ensures that data is up-to-date and enables changes to be logged.

Over the last 15 years, the Department of Clinical Airway Research of Fraunhofer ITEM has developed significant knowledge in the performance of clinical trials (mostly early clinical trials in phase I and phase IIa) under GCP (Good Clinical Practice). Electronic data capture systems and a quality assurance system including computer validation has been established which is a prerequisite for the acceptance of this work by regulatory authorities. A highly-qualified Proof-of-concept centre for early clinical trials, the Clinical Research Centre (CRC) Hannover, will be available to demonstrate early clinical trial test cases for the successful implementation of the BRICC Workbench and its impact on translational biomedical research outcomes (cf. Chapter 3).

Fraunhofer ITEM participated in the IMI project U-BIOPRED, evaluating new biomarkers of severe asthma in an unbiased systems biology approach. For that purpose, clinical samples were collected in different study sites in Europe under a consolidated quality regime. The samples were sent to the centralized Biobank in Manchester (UK Biobank). Individual samples were subsequently delivered out of the Biobank to different laboratories for end-point measurements. TranSMART has been used for central data sharing and hypothesis generation ((13); cf. Fig. 5).

If the BRICC Workbench had already been established with U-BIOPRED, a centralized yet virtual sample repository (metabiobank) would have spared time, effort and shipping costs, improved sample accessibility and transparency of research, without jeopardizing sample and workflow quality. Moreover, integration of the U-BIOPRED Biobank metadata into international IT infrastructure, as already implemented with p-BioSPRE (see above) and foreseen for HUB (see below), would leverage re-purposing and secondary use of these high-quality specimens and data.

Figure 5: U-BIOPRED Handprint
For the U-BIOPRED project, specimens and data from 17 study centres were collected and analysed over a tranSMART data management platform to form the final U-BIOPRED Handprint helping "researchers (to) better understand the different subtypes of severe asthma to offer better and more personalised treatment to those affected." (13)

Hannover Unified Biobank (HUB) is the central biobank of the Hannover Medical School (MHH). The biobank is located in the CRC building operated by Fraunhofer ITEM. Besides offering high-quality validated pre-analytic processes and storage of biomaterials, HUB is in the process of establishing a professional biobank information management system (BIMS) called CentraXX (Kairos, Bochum, Germany). The CentraXX system will be validated during this project (2018) to obtain the level of software quality needed for long-term Biobank storage. Thereafter biomaterials from clinical trials and associated study participant data will be collected and stored by HUB in close collaboration with Fraunhofer ITEM as a sample and data owner. This biobanking infrastructure will provide a proof of concept for the BRICC Workbench quality regime (cf. Chapter 2.3).

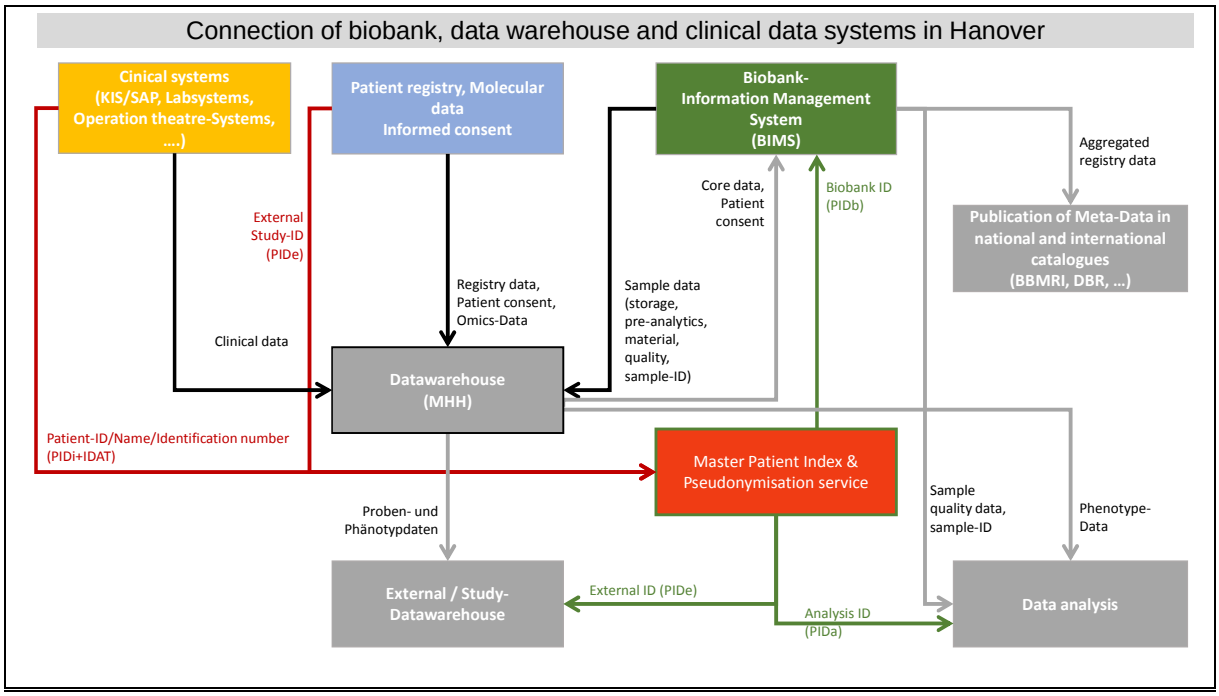

Figure 6: Data transfer and data protection schema at HUB
Patient registry (HIS /SAP), clinical data systems, biobank data system and data warehouse (DW) are tightly interconnected. Strict data security guidelines (pseudonymisation concept) are guaranteed. Meta data of registries are published in national and international catalogues (BBMRI, DBR).

To integrate existing and upcoming services (including the ones described above) and generate added value, a Cloud-oriented Service Market Place (CSMP; (14)) concept was proposed by Fraunhofer a few years ago. Featuring the flexibility and scalability of an open system, which might in the mid-term also support assembly of redundant / competitive services, we deem this concept well-suited to enable evolution of a truly comprehensive and sustainable biomedical research infrastructure such as an implementation of the BRICC Workbench. Although researchers and companies currently voice some scepticism and concern about outsourcing sensitive information to external servers, numerous technical solutions for secure cloud computing are well-advanced (15) and are expected to be available for the Workbench.

While, in contrast, technical solutions presented by the "Trusted Cloud" initiative ((16); recently successfully completed by the German Federal Ministry for Economic Affairs) apply to rather limited sectors of biomedical research and healthcare, a comprehensive legal framework (17) has been published by the working groups involved and will be delivered in time for the set-up of the BRICC Workbench.

As mentioned earlier, quality of services, data and specimens is of paramount importance for upcoming research in translational medicine. Since ISO (the International Organization for Standardization) launched their Technical Committee on Biotechnology (ISO TC 276; (18)), including a Working Group „Biobanks and Bioresources", at the end of 2013, we deem our initiative just in time to set up the BRICC Workbench's ICT and governance infrastructure in line with the same upcoming ISO standards that will be adopted by professional biobanks and clinical research centres for certification/accreditation. We note that a consistent, coherent, and compulsory quality policy of the entire infrastructure will be a key success factor of the BRICC workbench to fulfill industrial research requirements.

To summarize, an exceptional network of expertise concerning quality-assured data and information management in Life Sciences already exists, which will be used to support a timely development and implementation of the Workbench. It is expected that the core systems as developed for CRIP-based metabiobanks and in the ND4BB-TRANSLOCATION project can be used directly from scratch. New adaptations and developments have already been planned and the core systems natively designed to handle future enhancements, whether these are new functionalities or scale up/scale out options.

Beyond joining efforts, the three institutes involved (Fraunhofer IME, ITEM and IZI) will naturally draw upon the wealth of expertise provided by Fraunhofer as a whole, including specific groups and alliances (e.g. Fraunhofer Big Data Alliance), once the concept for the BRICC Workbench which is outlined in the following chapters has been finalized amongst project participants and stakeholders.

2. BRICC Workbench Concept

2.1 General Concept

The ultimate goal of the BRICC Workbench is to provide hassle-free, seamless access to biomedical research software tools and data without the inherent technical and legal complexities involved. Thus, over the Workbench, Fraunhofer intends to speed-up the evolution of the information ecosystem - in other words: the infrastructure - required for translational biomedical research (Figure 7). The latter being usually organized in projects of limited duration and resources, many independent solutions have been developed across the entire scientific community, in conjunction with various Fraunhofer institutes (cf. Chapter 1.). With the BRICC Workbench, we aim to overcome their fragmentation between these solutions and integrate them as crystallization nuclei to support an enduring, yet flexible biomedical research infrastructure.

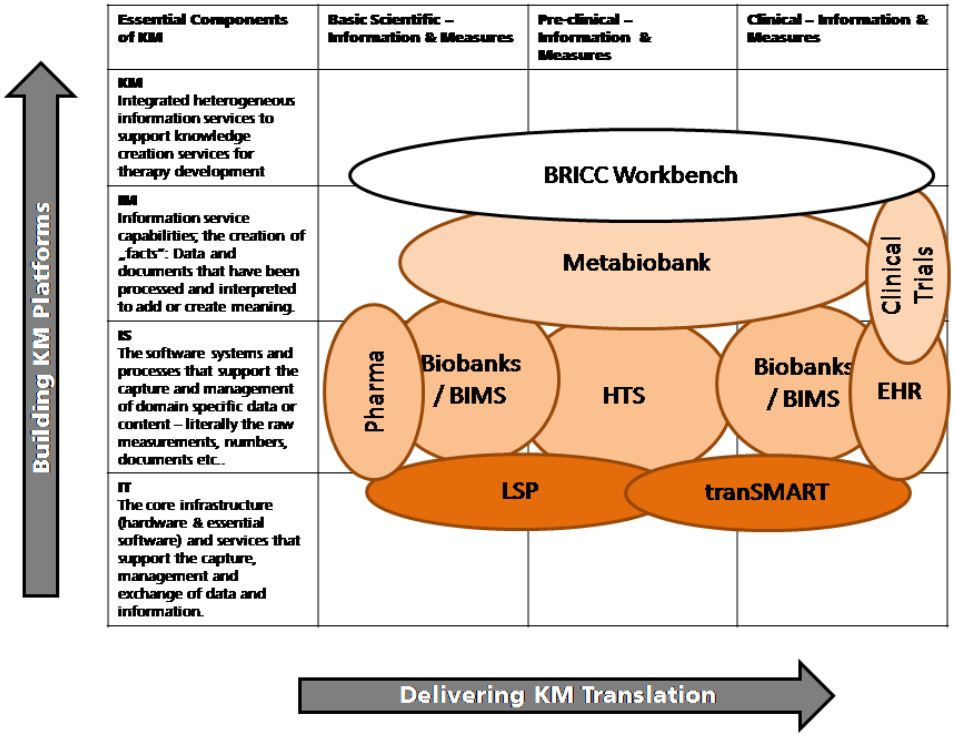

Figure 7: Knowledge Management Strategy / Matrix as envisioned by IMI (in black) and tools / modules (colored) which already exist with Fraunhofer, and which are envisaged for integration with the BRICC workbench.

In the past several years numerous initiatives have begun to functionally integrate data for biomedical research, a number of which have been described by Baker (19). However with these systems a substantial "attrition rate" is seen: Only a small proportion still remain under development, and considerably less have been implemented. Why do so many of these systems fail? And why do we believe that we can do better?

In our view, there are two main points to be considered; first, the rationale for data sharing: It might, for instance, be motivated by the need to establish statistical rigor for upcoming projects, and / or by efficiently performing in-depth data analysis, while having to cope with high levels of data complexity (20) anyway. What is missing is a comprehensive, user-friendly solution which can be used for all data sharing motivations from the very start, and which will continue to subsequently evolve. The proposed solution is inspired by a combined data integration process, driven by IT infrastructure wherever possible, and supported by specific data analysts where neces-

sary and appropriate. It will start out from a pre-existing, yet scalable and up-to-date metabiobank system, emalgamating data from two biobank networks on colon and prostate cancer, and enable data processing over a tranSMART-based system. The initial system will be swiftly enhanced by further applications and use cases.

Second, although it is mainly discussed in terms of data formats and integration, a sound ethical, legal, and contractual framework is a critical underlying prerequisite for a viable data integration platform. The privacy- and confidentiality-preserving solutions that are available for the Workbench, which have already been developed and implemented with Fraunhofer and several pharmaceutical companies, are presented in Chapter 2.4.

This White Paper is an initial proposal presented to the BRICC Workbench's future customers and partners for discussion, design thinking, and improvement; so any comments and criticism are very much appreciated.

2.2 ICT Architecture

The initial architectural proposal for the BRICC Workbench is depicted in Figure 8. The Workbench is intended to offer many more tools, including open source software and applications from external software developers (cf. Chapter 2.4; "Content creators"),

The overall architecture of the Workbench follows a service-based approach, which will allow individual tools provide a set of external services for other components, while abstracting away the complexity within the individual pieces of software. The modular approach facilitates, by means of standard interfaces, coupling and decoupling of components as needed, e.g. for XML-files, RESTful webservices and the like. By the same token, this architecture guarantees scalability, since software components can be distributed across multiple servers, enabling load management strategies where applicable.

2.2.1 Service Layers

Actors from pharma, biotech and academia possess massive databases from which data must be extracted for further processing and analysis. The BRICC Workbench facilitates this process by providing the tools needed to extract, anonymize (if applicable), harmonize and transfer such data. In order for this "big data" to be usable in the wider context of the BRICC Workbench, it must be interpreted by a ***semantic integration layer***, which shall encapsulate the complexity of reconciling the semantics of disparate data sources by means of a ***metadata layer*** and pertaining dictionaries and catalogues.

Once data has been appropriately harmonized, it must be fed into other tools, which will provide the actual customer services, such as biospecimen locators or metabiobanks, omics data browsers, biomedical data warehouses (Cf. Chapter 2.2.2). This is accomplished via an ***ETL layer***, which provides tool-specific interfaces to various systems, both open source and Fraunhofer-proprietary.

The core services needed by the BRICC Workbench are based on components of the Fraunhofer CRIP Toolbox (10), an industry-validated software suite for metabiobanks, which will provide a major head start into the Workbench's development. The CRIP Toolbox features a number of versatile modular services which address different needs in metabiobanking, e.g. integration and presentation of highly variable data from different data sources, such as classification of diseases in raw text medical records or presentation of metadata in a search tool, high volumes of omics-data, fast access to biomaterial and associated data records.

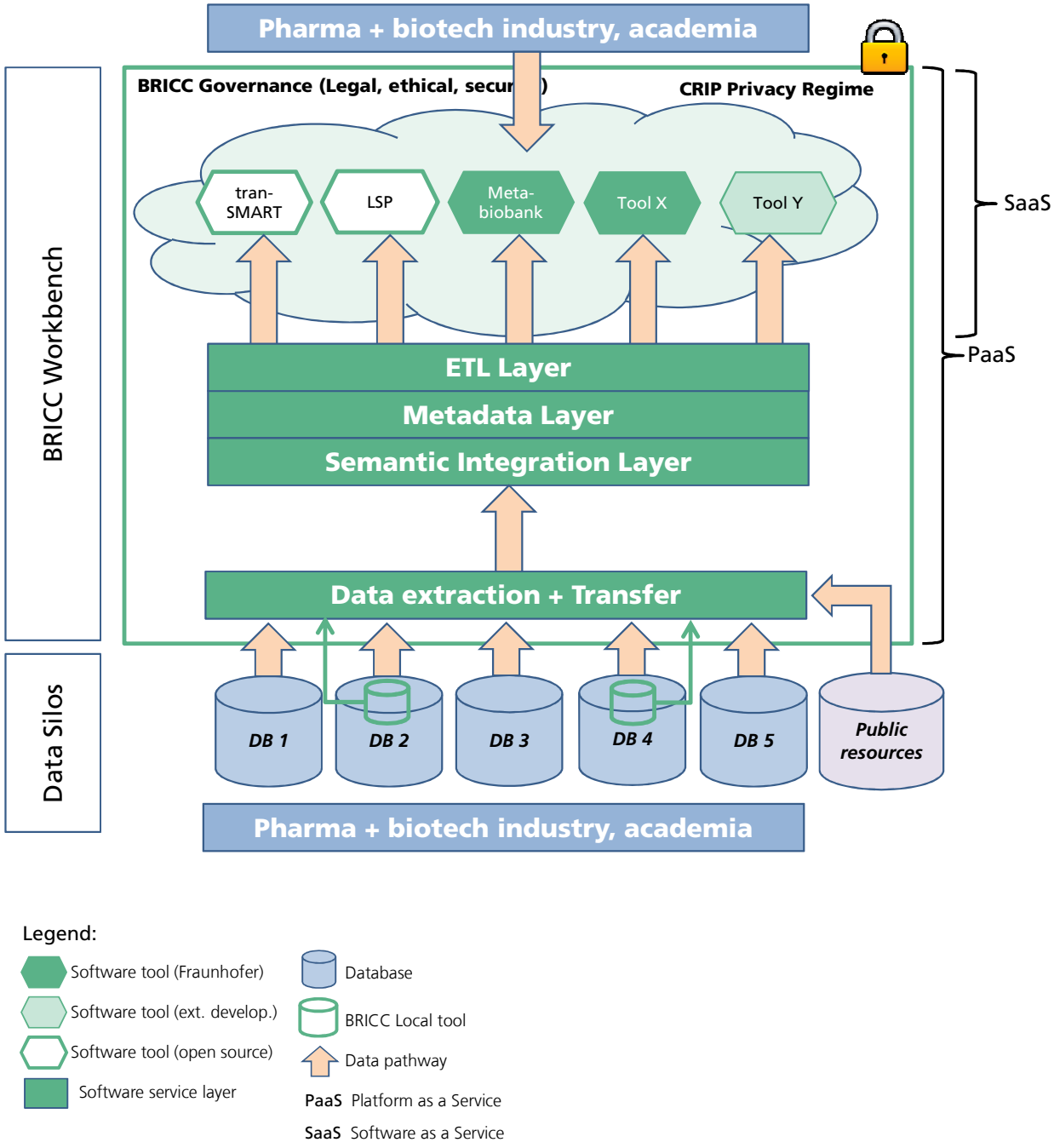

Figure 8: Overall Architecture of Fraunhofer BRICC Workbench

The aforementioned complexity involved in reconciling disparate data sources is handled by CRIP.IDB or CRIP.IANUS tools. Biobank operators can import their (pseudonymized) data in different formats (.csv, .xls, CDSIC ODM, etc.), which are then pre-processed on-site (cf. "BRICC local tools" in Fig. 8) using biobank wrappers, and harmonized through a metadata repository (MDR), followed by upload to the central database (CRIP.CDB). To ensure compliance with privacy regulations, prior to upload, data are anonymized by the CRIP.Anon module, which employs de-identification and/or k-anonymization techniques on a configurable basis. To enable responsive query times regardless of database size, CRIP.Index creates a live index of all available

bio-specimens in different CDB pools along with parameterized query combinations, allowing swift retrieval of desired specimens.

2.2.2 Software Tools

BRICC Workbench's bold ambition is to set up a "one-stop shop" for a wide variety of user scenarios, catering to the needs of customers in pharma, biotech and academia. It will provide a full suite of solutions, comprising of infrastructure availability, data extraction, transformation, standardization and harmonization. Upon this overarching foundation – the BRICC cloud – specialized open-source biomedical tools, such as tranSMART, LSP, etc. can be instantiated, according to concrete user needs. Besides such publicly available tools, the BRICC Cloud also provides the necessary ICT infrastructure to run Fraunhofer-proprietary tools such as the CRIP Search Tool, and an XML-based web interface dynamically generated according to the metadata information stored in the CDB or CRIP.Index. Communication between the web services and the actual data repository is achieved deploying the software interface CRIP.API, enabling researchers to not only query desired biospecimens but also to forward project requests directly to relevant biobank partners.

The tools shall be provided under two major models, both as Platform as a Service (PaaS), as well as Software as a Service (SaaS); cf.Fig. 8. In the first model, vendors and external software developers (content creators) can take advantage of the available platform services to host their own tools, either component-wise or platform-wise. This open approach enables the establishment of a ***BRICC cloud marketplace***, where vendors can trade BRICC-based software solutions with other users. As we shall see, this represents one of the cornerstones of BRICC's business model (cf. Chapter 4). In the second approach, the BRICC Workbench itself will provide out-of-the box, hosted solutions for customers, who will not need to worry about time-consuming issues such as data harmonization and system availability.

2.2.3 Data Curation and Collaborative Access (ColA)

The most valuable item of the BRICC Workbench is the data provided by partners. Thus, it is essential that the proposed solution focuses on standardized terminology and preserving the uniqueness of the data sets.

As of February 2015, biosharing (18) have identified 69 reporting guidelines, 336 terminology artefacts, and 174 exchange formats in the field of bioscience. This mass of terminologies and formats cannot be handled by the BRICC Workbench *ab initio*. In addition, nobody in the lab can possibly be expected to know the differences or pros and cons of this plethora of standards. Obviously this is the field of experts. And where are these experts? Not where the data is generated. These experts are based alongside managed repositories, as it is vital that data is harmonized before it is uploaded to these types of repositories.

An optional way out of this dilemma is the creation of a service for data curation, performing correct translations, not only of terms into predefined values, but also to suit the purpose of cross study evaluation. Our goal is a data integration process driven by data analysts (and not only by IT infrastructure) that can evaluate previously unrelated data from isolated studies to generate knowledge. Themain focus is on comparability of the data and flexible, user driven evaluation.

This will be achieved by combining tools (open source and commercial) developed for specific data collection and evaluation tasks, and adding well-established and widely-accepted frontend tools and services for scientific data management to improve the value of data.

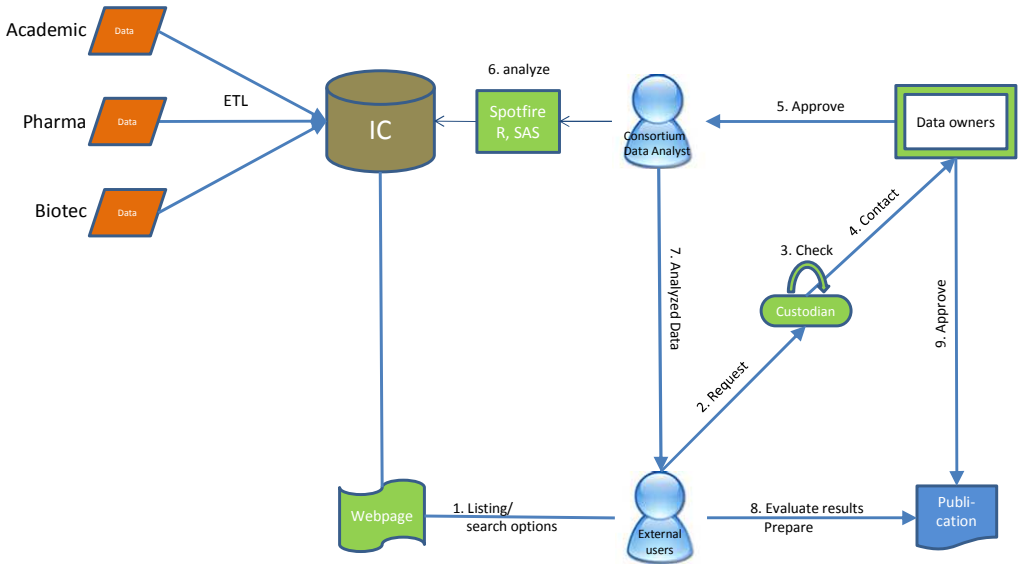

Figure 9: Collaborate Access Framework (ColA)

External users need to approach data in the repository via a custodian (Figure 9). This provides public access to the valuable data in the repository while IP and data security is maintained by trained, internal data analysts.

2.3 Quality Regime

Quality Management and Quality Assessment (QM & QA) for the BRICC Workbench's entire workflow will be a key success factor. Hence QM & QA need to be implemented at the following points of interaction (cf. Chapter 2.4):

- Sellers (Service providers, e.g. biobanks, sequencing, HTS, bioanalytics)
- Content creators (Software developers, hospitals, data providers)
- The Platform (Cloud?) as such

Accordingly, different quality standards will generally apply throughout different tasks or projects processed over the Workbench: Starting for instance with biobank data, biobank quality standards (see below, and cf. Chapter 1) will apply first of all. Second, the biobank information management system (BIMS) might need to be validated according to Good Clinical Practice (GCP) regulations if biospecimens are meant to be preserved from clinical trial patients, or, following generation for diagnostics or surgery, subsequently annotated with clinical information for research, e.g. for early biomarker discovery. Third, data have to be integrated on a certified platform to ensure high performance, usability, and accountability for queries, project requests and agreements.

As another example, starting out from high-throughput analytical phenotyping based on Good Laboratory Practice (GLP) or GCLP (21), a standardized Laboratory Information Management System (LIMS) might be required, the output of which is then fed into a quality-controlled data warehouse and processed by accredited data analysts. Again, results will have to be conveyed to customers over a certified platform, including accounting services.

Biobank quality has been discussed for long time as a major bottleneck for the identification and validation of new targets (3) for therapy and diagnostics. To overcome these problems and

make improvements to pre-analytical biospecimen quality, tools like SPREC (Standard PREanalytical Code; (22)) or BRISQ (Biospecimen Reporting for Improved Study Quality; (23)) have been developed and can easily be integrated in the BRICC Workbench upon command by the user. The Technical Specifications filed by the SPIDIA consortium (24) for extraction of nucleic acids and proteins from blood and tissue are currently in the process of being converted into CEN and ISO standards which will certainly be adopted by professional biobanks and mirrored in Fraunhofer metabiobanks.

To improve biobank quality, a multitude of guidelines and best practices have been published (25) over the last couple of decades. At present these numerous approaches are being collated into binding global standards by the Working Group "Biobanks and Bioresources" of the ISO TC 276 Biotechnology (26) and are expected to be available by the end of 2017 (Figure 10). This endeavor will eventually lay the ground for global interoperability and comparability of biospecimens and finally provide a sound basis for a consistent and comprehensive, yet binding standards-based quality regime across the BRICC Workbench.

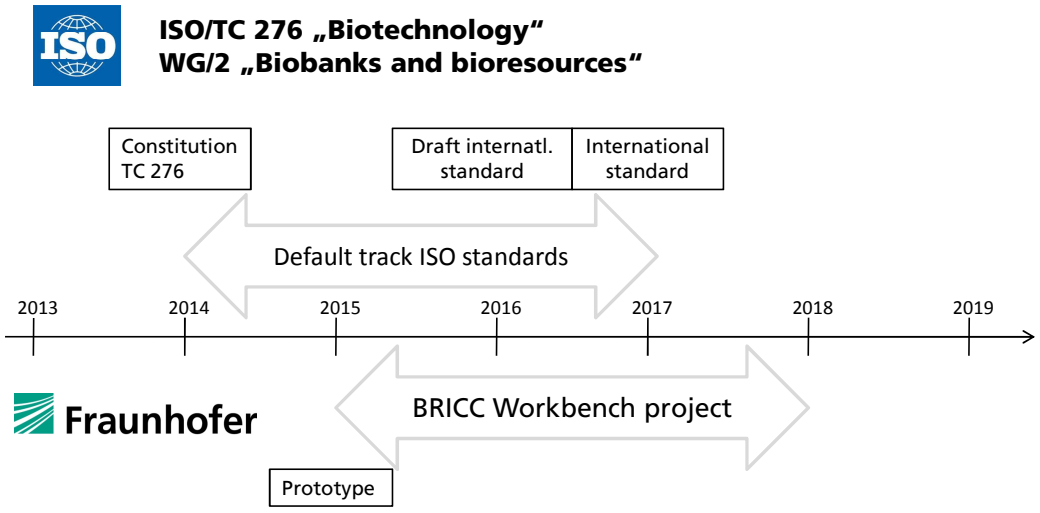

Figure 10: Timeline development of ISO standards for biobanking vs. set-up of BRICC Workbench

Before start-up and during further development, any components marked for integration into the Workbench need to undergo a benchmarking procedure, for which a framework and policy will be established (see Fig. 10 and below).

In its entirety, the Workbench will inevitably mirror a patchwork of quality standards which therefore need to be included in datasets on all levels of the Workbench's workflow. Integrating all these quality parameters into the Workbench's tools and services will allow us to set up an outstanding quality framework delivering on the promise that "...all services and processes available over the Workbench will ... support users' subsequent submissions for marketing authorization" for their products.

2.4 Governance Structure

From its conception, the BRICC Workbench will cater to three primary customer sets (Fig. 11):

- Consumers (Pharmaceutical and Diagnostics Industry)
- Sellers (Service providers, e.g. biobanks, sequencing, HTS, bioanalytics)
- Content creators (Software developers, hospitals, data providers)

The BRICC Workbench is a patient-centric endeavor (cf. Fig. 11). However patients will not directly interact with the Workbench, but rather with one or more segments of its customers (e.g. hospitals, biobanks, pharmacy and diagnostics).

From its very start, the Workbench's management will be able to draw upon a pre-existing legal and contractual framework and expertise developed for CRIP and IME-SP's InfoCentre, including

- a Biobank Data Transfer Agreement, mirroring the privacy-preserving CRIP concept for metabiobanks (11), and since 2006 approved by German data protection authorities. This contract has, moreover, been approved by a dozen of legal departments of universities / medical schools / hospitals, by the Fraunhofer legal department, and by the Institute of Legal Informatics (27) at Hannover University for biobank access within the EU-project p-medicine (28). The contract has been closed between Fraunhofer and a dozen of legal departments of universities/medical schools / hospitals.
- a Model Contract for Biobank-based projects (approved by several pharmaceutical industry legal departments)
- a Data Transfer Agreement (in preparation by ND4BB-TRANSLOCATION project in collaboration with lawyers from diverse partners and Fraunhofer as data hosting provider).

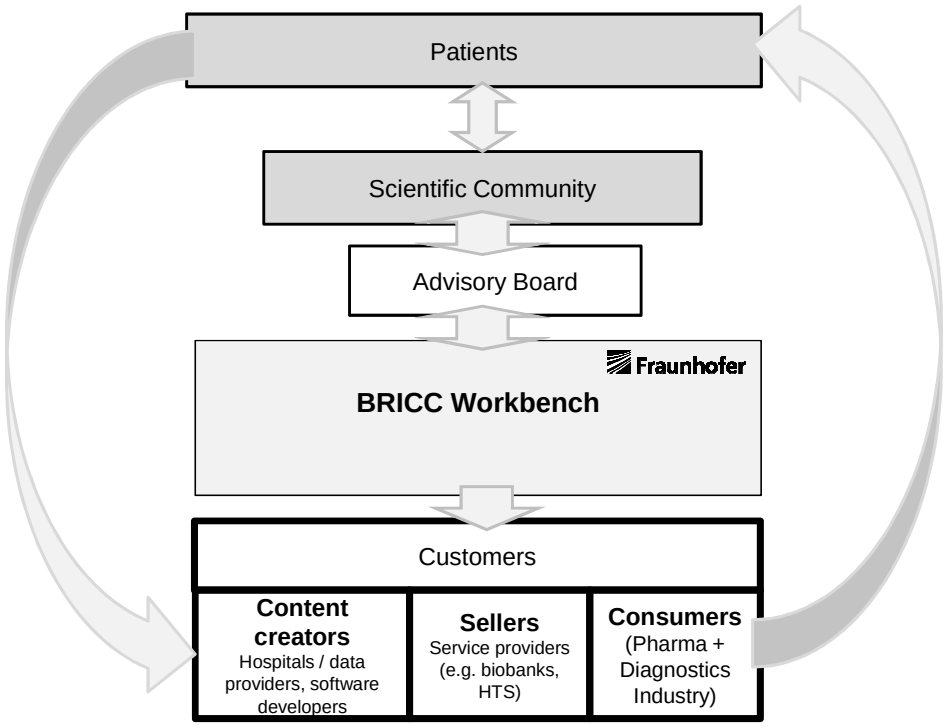

Figure 11: Overview over the governance structure of the BRICC workbench

The primary management and governance structure for the BRICC Workbench (light grey box in Figure 11) is expected to be developed in detail and set up by a joint project (2016 - 2018) along the lines of a Cloud-oriented Service Market Place (CSMP), in order to efficiently support its subsequent operational phase. During the project phase (2016 - 2018), and if applicable the Fraunhofer MaVo funding scheme (cf. 4.4) would for instance prescribe that, in addition to the project management, two external project tutors ("Projektbegleiter") be involved by subcontracting, and that the project is overseen by the Board of Directors of the participating Fraunhofer institutes.

Moreover, broader interdisciplinary expertise will be required to secure full compliance of the Workbench with all ethical and social requirements, as well as appropriate representation of the project's stakeholders. Hence an independent interdisciplinary Advisory Board will be responsible to file strictly patient-centric policies for all quality issues as well as ethical, legal, and social implications of the research that will be mediated and conducted over the Workbench. The Advisory Board will also help secure continuity of governance and sustainability beyond the project phase.

The CRIP Advisory Board (29) who have been active according to their Rules (30) from 2006 - 2015 provide a successful blueprint for the BRICC Workbench Advisory Board. In particular, the CRIP Board's rule to have the different partners and stakeholders (cf. Fig. 10) participate in the Board meetings as invited guests has proven to be a flexible, comprehensive, and equitable means to involve the project's entire spectrum of stakeholders - eventually right up to patients - into the infrastructure's progress.

3. User Scenarios

The following shall provide a non-exhaustive list of some user scenarios that could be supported by the BRICC Workbench.

3.1 Biobank Access: From Metabiobanks towards Biobanking on Demand

In 2015, Fraunhofer IZI-BB's existing metabiobanks (or biospecimen locators) CRIP (4) and P2B2 (31) are undergoing a thorough technical re-design and organizational consolidation into a joint metabiobank portal. Details of this re-structuring (including the name of the new metabiobank) will be discussed with CRIP Advisory Board and biobank partners in a Board meeting by the end of April, and subsequently implemented. By aiming to become a crystallization nucleus of the BRICC Workbench, the new metabiobank will feature utmost flexibility for biobank partners and their deep datasets, and the quality regime set out for the BRICC Workbench (cf. Chapter 2.3), including certification of critical metabiobank software modules and workflow.

As described in Chapter 1, p-BioSPRE, is a state-of-the-art metabiobank that has already been introduced. Browsing the p-BioSPRE query interface, a user will stratify required cases ("strata"; cf. Fig. 1) and specimens and retrieve "pools" of cases that match his/her request. Upon submitting a request ("project") for such pool of cases, the set of requirements/parameters entered by the user is distributed on-line to the participating biobanks. Conversion of metabiobanks into sophisticated prospective project request systems ("Biobanking on Demand") is already anticipated by this workflow. Instead of administrating so-called "bio-vaults" (as lately discussed at ISBER and ESBB meetings), such systems would trigger the prospective collection of samples and consent actually required for research, and subsequent sample preservation across a network of competent accredited biobanks. Few advanced biobanks' BIMS nowadays already provide an interface e.g., to their Institute for Laboratory Medicine or Department of Surgery, that allows them to preselect patients from whom informed consent should be collected and remnant samples should be preserved.

Moreover, the BRICC workbench ***itself*** can be used to extract knowledge by featuring, for instance, query tools tailored to industry needs (Table 1). Taking advantage of the data privacy framework provided by the BRICC workbench and upon agreement with the relevant partners, information from biobanks registries could be queried transversally across partners. Such infor-

mation shall be collected by means of biobank questionnaires specifically tailored to the industrial needs (cf. Table 1). To ensure international interoperability, the use of standards shall be pursued, for example utilizing approaches such as BBMRI's MIABIS (32).

Table 1: Exemplary excerpt from an industry-oriented biobank questionnaire
(Courtesy Ann Cooreman, Tissue Solutions Ltd., and Dr. Arndt Schmitz, Bayer Pharma AG)

Question	***Answer options***
Sample characteristics	
Processing time known? from warm ischemia to preservation	YES/NO
Matched samples available?	YES/NO
Multiple time points?	YES/NO
Can patients be asked for further samples?	YES/NO
Custom prospective collections possible?	YES/NO
Ethics and Consent	
Are there exclusions in your ethics permission? (Indicate for majority of top	One database field with predefined options – NO, genetics, genomics, cell lines, animal
Are consent templates available in English?	YES/NO
Are consent templates on your homepage?	YES/NO
Donor Data	
Do you include in basic data gender, age, BMI, ICD indication?	Yes, electronically; Yes, paper; No YES/NO
Are extended clinically data available?	Yes, electronically; Yes, paper; No
Functional annotations and measurements available (e.g. ECG, ultrasound)?	Yes, electronically; Yes, paper; No
Imaging results available (CT scans results, MRI results)?	Yes, electronically; Yes, paper; No
Extended analytical clinical chemistry laboratory data available?	Yes, electronically; Yes, paper; No
Genetic data available (ONC mutations)?	Yes, electronically; Yes, paper; No
Therapy / prescription data available?	Yes, electronically; Yes, paper; No
Therapy outcome data available?	Yes, electronically; Yes, paper; No
Survival data available?	Yes, electronically; Yes, paper; No
Process Info	
Do you want info on samples back?	Yes, any raw data; yes, aggregated data; publication mandatory; publication desired;

3.2 tranSMART integration

The open source tranSMART system is gaining a stronger foothold in industry as the go-to biomedical data warehouse solution, and it will be an integral component of the BRICC Workbench. It shall cover the analytical and bioinformatics aspects necessary for full utilization of the available data, bringing together clinical trial, metadata, gene profiling, and biomarker data. Despite its increasing popularity, the tool's deployment and continuous operation remains a formidable challenge even for seasoned (bio)informaticians, and often requires the commission-

ing of external service providers for this task, which is, in any scenario, significantly time and resource-consuming. On the top of that, curating, harmonizing and feeding data into tranSMART can be even more daunting and challenging, in spite of the available ETL (extract, transform, load) procedures available in the community. The BRICC Workbench shall abstract away the inherent complexity for the use of this tool by making on-demand tranSMART instances available out-of-the box, which are fully-integrated with the underlying BRICC ICT infrastructure, providing seamless, hassle-free data integration.

Case in point: a given clinic utilizes the biobanking BRICC Workbench capabilities to make biospecimens available for research. The biospecimens were collected in different clinical studies within which, e.g., a survival analysis must be carried out. tranSMART offers ready-to-use, one-click algorithms for survival analysis. For that, instead of going through the hassle of having to set up a tranSMART instance, develop and validate the necessary ETL to get the data into the tool, the clinic can simply "activate" a tranSMART instance on the BRICC workbench. Since the necessary database and server infrastructure as well as the (BRICC-provided) harmonization and ETL layers are already in place, data from the institution's clinical trials are seamlessly copied over the respective datamarts in tranSMART, enabling direct utilization. In addition, data synchronization services in the BRICC Workbench guarantee the actuality and accurateness of data.

3.3 Fraunhofer ivD Platform

The Fraunhofer ivD Platform, a Point-of-Care in-vitro diagnostics platform, provides parallel detection of up to 500 biomarkers, utilizing both immunoassays and DNA-based tests, and offers results in minutes. The wealth of data generated by such a platform can alone provide invaluable insights for clinical practice and research. With a convenient, minimally invasive procedure for biomarker measurements, treatment responses can be more closely monitored and adverse effects detected quickly with minimal discomfort for patients. Furthermore, taking advantage of the tranSMART interfaces, (raw) microarray data from the platform can also be integrated into the BRICC workbench.

Notwithstanding, in combination with other data sources such as those harbored by the BRICC Workbench, the potential of Fraunhofer ivD platform's data for translational medicine cannot be overstated. In a conceivable scenario, gene expression data that is already available could, e.g., be effortlessly matched with rich biomarker measurements, since they are under the same ICT infrastructure (BRICC). This could possibly enable the discovery of novel relationships and correlations and/or reduce the bench-to-bed duration for innovations. For this reason, biomarker measurements originating in the Fraunhofer ivD platform will be directly integrated into the BRICC ecosystem of tools.

3.4 Pharma and biotech industry

The development of novel drugs is largely driven by data coming from multiple silos or "small data" (clinical trials, gene expression profiling, pathways, biomarkers, biobanks, simulations, omics data, etc.) that must be analyzed as a coherent whole ("big data") if actionable insights are to be obtained. While tools exist which aim to provide pharma companies with an easy way to tap into their data silos such as open-source solutions (e.g. tranSMART) or proprietary ones (Thomson Reuter's Cortellis and Pipeline Data Integrator), data harmonization remains a challenge. As previously discussed, one of the core components in the BRICC workbench is a comprehensive semantic layer aimed at taming the complexity involved into unifying disparate data sources. Therefore, in addition to facilitating out-of-the-box access to otherwise resource-demanding tools such as tranSMART, pharma companies will be able to use BRICC workbench's capabilities/services to unify their data silos wherever they may reside.

3.5 Big data analytics

While handling individual data silos and bringing them together is already challenging enough, analysis of the resulting large datasets bears great potential, especially considering the long-term growth of the platform as whole. As the BRICC workbench grows organically, what new interrelationships can be established which were not previously known? How do biomarkers measurements – across studies – evolve over time? How does disease history affect drug response? To which extent can patient response to a drug be predicted?

To be able to fully utilize this opportunity, modern big data analytics approaches must be employed in order to establish associations and correlations to identify clusters and perform prediction / forecasting. The big data vision for the BRICC workbench is long-term, but to make this possible in the future, BRICC workbench will be designed in such a way as to be "big-data friendly", meaning that it will be possible to extend it to apply big data tools such as R, Hadoop, Spark, MapReduce and the like to mine the data and drive new insights. The same applies for visualization and data mining tools such as IBM Many Eyes, D3.js, Weka, Raphaël and others.

Furthermore, the cumulative know-how and expertise in the field that will be acquired by the BRICC workbench technical team will make it possible to position itself as an all-round platform and service provider for intelligent analysis of biomedical big data.

4. Business Model

While it is not possible, at this stage, to exhaustively and definitely establish BRICC's business model, three possible major drivers of revenue/growth largely utilized by industry can already be identified.

4.1 Resource subscriptions

In alignment with the widely employed SaaS model, the BRICC workbench will charge its users a fee, either monthly or service-based, according to which tool access type is desired. New services can be purchased as needed, without up-front costs. Users are thus granted access on a flexible pay-per-use basis. In a similar fashion, users can not only purchase access to existing hosted tools but also book platform-wide resources (such as servers, databases, connectors) in order to host their own services. Well-established cloud computing platforms, such as Microsoft Azure (33) utilize such a subscription model.

4.2 BRICC market place

By taking advantage of the service APIs of the BRICC Workbench and the platform resources, "content creators" can develop solutions, which might be of interest for other users in the BRICC community. Such solutions can be traded with other users and "plugged" into existing subscriptions to, e.g., facilitate routine operations such as big data processing using Hadoop or MapReduce, thereby creating a BRICC market place. A nominal platform fee shall be levied upon each transaction. A similar approach is utilized by the Salesforce CRM (34).

4.3 Technical consultancy

While BRICC Workbench's ultimate goal is to provide hassle-free, seamless access to biomedical software tools without any of the inherent technical complexities involved, specific scenarios will nevertheless require a certain level of ICT expertise. Users can therefore – if so requested – con-

sult BRICC experts to assist in setting up and configuring the tools required, as well as eliciting performance and user requirements.

4.4 Seed financing

To kick the BRICC Workbench off, the consortium of Fraunhofer IZI, IME and ITEM is currently seeking support from private companies and funding agencies. One of the consortium's options would be to apply for Fraunhofer MaVo funding. With this scheme, Fraunhofer might support "Marktvorlaufsvorschung" for three years - assuming that a certain amount of the funds will be reached as revenues from industry within one year after termination of the funding phase.

4.5 Sustainability

A business model for sustainable operation of the BRICC Workbench is to be developed during the project phase after diligent market analysis.

5. References

(1) http://clinicalcenter.nih.gov/ccc/btb/
(2) ESFRI; insbes. BBMRI: http://www.bbmri-eric.eu
(3) Prinz, F., T. Schlange, et al. (2011). "Believe it or not: how much can we rely on published data on potential drug targets?" Nature reviews Drug discovery ***10***(9): 712-712.
(4) http://crip.fraunhofer.de
(5) http://www.imi.europa.eu/
(6) http://www.m4.de/
(7) http://transmartfoundation.org/
(8) Weiler, G., et al. (2014). *p-BioSPRE – an information and communication technology framework for transnational biomaterial sharing and access*. ecancer 2014, 8:401
(9) http://www.crip.fraunhofer.de/en/toolbox
(10) Schröder C, Heidtke, KR, Zacherl N, Zatloukal K, Taupitz J (2010). Safeguarding donors' personal rights and biobank autonomy in biobank networks: the CRIP privacy regime. Cell Tissue Bank 12(3): 233 – 240
(11) http://stratumbiobanking.org/docs/STRATUM%20Data%20and%20Register%20Final%20Report%20v1%200.pdf
(12) http://www.ribolution.org/
(13) http://www.europeanlung.org/en/projects-and-research/projects/u-biopred/home
(14) http://www.cloud.fraunhofer.de/content/dam/allianzcloud/de/documents/ISST_CSMP-Whitepaperwwwfinaltcm421-98473.pdf
(15) http://www.cloud.fraunhofer.de/
(16) http://www.trusted-cloud.de/
(17) http://www.trusted-cloud.de/369.php
(18) http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=4514241
(19) Baker M (2012). Quantitative data: learning to share, Nature Meth 9 (1), 39 - 41.
(20) http://www.biosharing.org/
(21) Sarzotti-Kelsoe, M., J. Cox, et al. (2009). "Evaluation and Recommendations on Good Clinical Laboratory Practice Guidelines for Phase I–III Clinical Trials." PLOS Medicine 6(5)
(22) Betsou, F., et al., and ISBER Working Group on Biospecimen Science (2010). "Standard Preanalytical Coding for Biospecimens: Defining the Sample PREanalytical Code." Cancer Epidemiol Biomarkers Prev; 19(4) April 2010; DOI 10.1158/1055-9965.EPI-09-1268
(23) Moore, H. M., A. Kelly, et al. (2011). "Biospecimen Reporting for Improved Study Quality." Biopreservation and Biobanking ***9***(1): 57-70; DOI 10.1002/cncy.20147
(24) http://www.spidia.eu/
(25) http://www.p3gobservatory.org/repository/guidelines.htm; accessed February 24, 2015
(26) http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=4514241; accessed February 24, 2015
(27) http://www.iri.uni-hannover.de/research-1906.html
(28) http://www.p-medicine.eu/
(29) http://www.crip.fraunhofer.de/en/about/advisoryboard
(30) http://www.crip.fraunhofer.de/htdocs/pdf/Ordnung_IBMT.pdf
(31) https://p2b2.fraunhofer.de/index.html
(32) http://online.liebertpub.com/doi/abs/10.1089/bio.2012.0003
(33) azure.microsoft.com
(34) www.salesforce.com