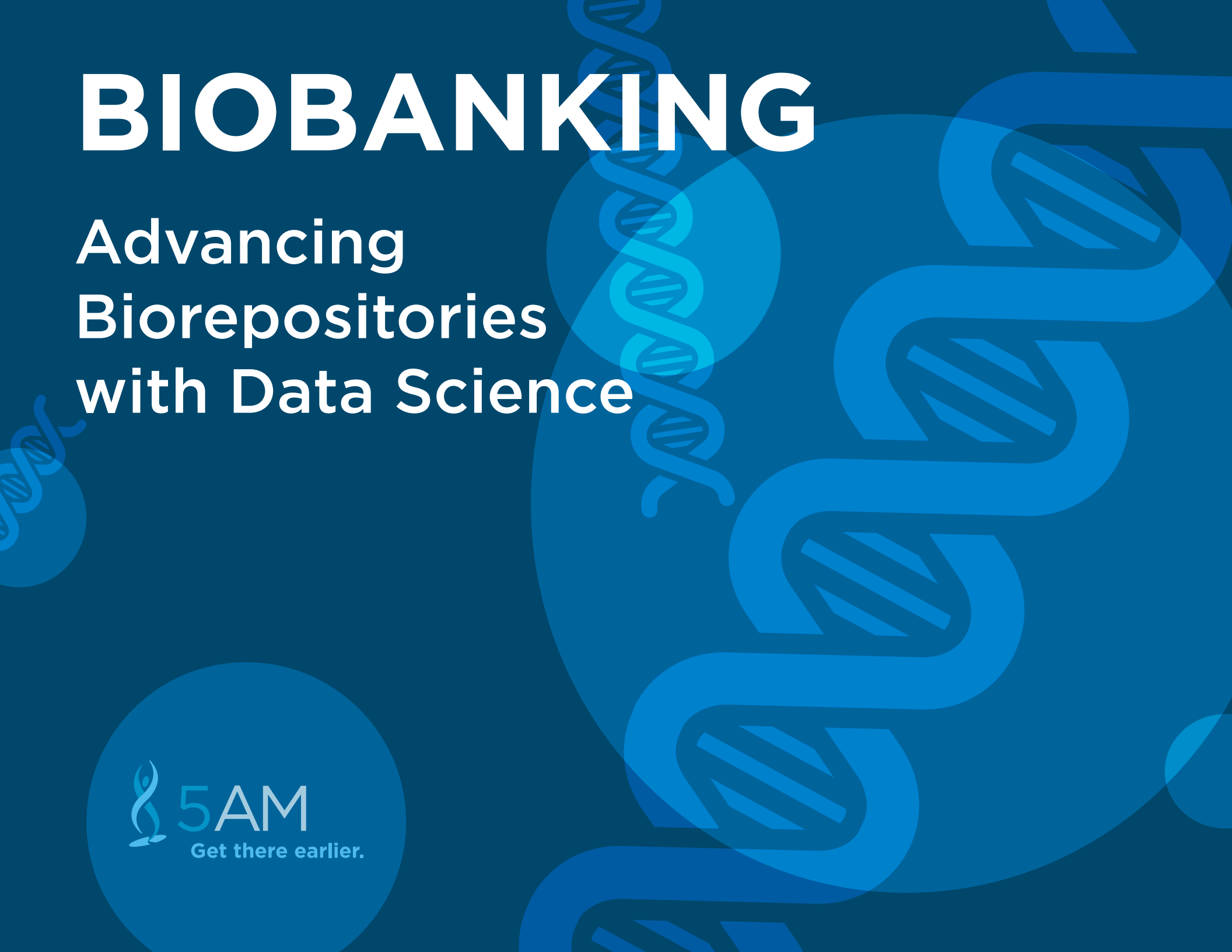
BIOBANKING

Advancing Biorepositories with Data Science



Contents

Overview

Biobanking has seen many changes over the past decade. Decentralized biobanks managed by spreadsheet have given way to institution-wide efforts that are managed through large scale information systems that can interoperate with laboratory information management systems (LIMS) and international databases that publish the resulting research. This trend is continuing through the use of tools like Biolocator to aggregate information about biospecimens from many institutions to allow researchers from around the world to build effective sample sizes for even some of the rarest diseases.

All of this is being driven through the adoption of data standards in these information systems. Data standards need to work not only on how data is expressed, such as data models, class names, and data formats, but also on what is meant by the data. Clearly expressing the semantic components of the data allows people and systems to interpret it without consulting the originator.

In this book we will discuss the issues around data standards and interoperability in biobanking, including existing best practices and gaps in current practice, the use of controlled vocabularies, how semantics improves the effective power of biobanks, different ways of expressing those semantics, and some thoughts about the complexities of identifying cell lines. Our aim is to inspire a discussion of future directions and set the stage for wider and more efficient use of biospecimens.

When it's time to scale to the enterprise and share accessing biospecimens and associated data, there are plenty of LIMS and biobanking vendors out there selling the software. However, the software product is only one piece of the puzzle.

Advancement of Biorepositories

Many institutions have biorepositories spread throughout multiple research centers or medical facilities. They are often found using differing solutions within each lab such as excel, homegrown databases or LIMS tools. Given the differing data elements and formats, data may be exchanged through email, thumb drives, or pushed to a folder on a local server.

When it's time to scale to the enterprise and share accessing biospecimens and associated data, there are plenty of LIMS and biobanking vendors out there selling the software. However, the software product is only one piece of the puzzle. In order to make it effective for the long run, and a reliable service to researchers, the institution must consider the following:

Business Workflow

Analyze the workflow between departments using **business process modeling** before diving into the product.

- How are biospecimens collected, stored, requested and distributed?
- Which activities are stakeholders involved in?
- What are the stakeholder's needs downstream with regards to biospecimens and associated data?

It is important to evaluate the as-is processes against the product's capabilities to define a to-be workflow that fits the overall purpose. When looking at the enterprise, the business processes for storing and accessing biospecimens need to align across multiple facilities or labs.

While designing a to be process for your institution takes precedence, some compromise is needed here from both the customer and vendor. Insisting on a specific business workflow can require expensive customization of the product and lengthy rollouts. A rigid workflow model from a product may require large business process changes and re-training across the board. Evaluate products that are aligned with good biobanking practices and are built to be **CFR Part 11** compliant.

Best Practices

The lack of high-quality biospecimens has been recognized as a blocker to research. The requestor has to trust the source biospecimen and its associated data to ensure quality research results. Biobanking best practices such as those from **NCI Best Practices for Biospecimen Resources** must be consistent

across the board to ensure quality of the specimens. This will require a review and governance of biobanking procedures (e.g. consent, collection, processing, storage, request, distribution) to incorporate the software as a part of the process, along with training.

Data Standardization and Data Sources

Accessing specimens and data from multiple sources starts with consistency across the biobanks. Global terminology standards relevant to biobanking include the **Common Biorepository Mode** from NCI, **LOINC**, **ICD-9**/ICD-10, and **SNOMED-CT**. If the institution has a data governance group or committee, it is best to align with this direction on vocabularies to ensure interoperability with existing and planned data sources. If not, have a plan to get there.

As the use of the biobanks increase, access to related clinical data is desired from various source systems (e.g. Pathology, Clinical Research Management Systems & Clinical Data Management Systems). This data must be mapped and related to the biospecimens. A data warehouse can act as the broker for mapping the multiple data sources to a set of standards for the enterprise.

Overwhelming? Don't panic, avoid analysis paralysis, and phase it in over time. Start small and allow plenty of time for piloting to adjust before rolling it out to the enterprise.

Why Data Standards Matter to Biobanking

A Biobank refers to any organized collection of biological material that once was either part of a living organism or produced by it. While this blog post focuses on the human biospecimen repositories, the fundamental principles discussed are relevant for many if not most other biobank types. Furthermore the terms specimen, biospecimen and sample are used interchangeably.

Biospecimen have long been a key asset in evidence-based medicine. In fact many 'lab values' in modern healthcare are derived from blood, urine and other biological samples. One can argue that whole medical disciplines, such as oncology, are founded on biospecimen-based diagnoses. Beyond their significance in patient-care, biospecimens also play a pivotal role in biomedical research. They help us understand disease mechanisms and develop of new molecular diagnostics and therapeutics.

The value of biobanks is not just determined by the quality of the banked specimen but also by the quality, richness, and representation of the information associated with these specimens.

> "Today, the samples are collected for tomorrow, therefore, improvement is needed now in standardization, automated enrichment of annotations from hospital information systems and disease registries, insight in overlapping collections of different forms of tissue banking and cooperation in national and international networks."
>
> —Biobanking for interdisciplinary clinical research

The Standards Deficit

There are billions of specimens held in biobanks worldwide. However their utility has been hampered by the absence of globally accepted biobanking standards, especially when it comes to appropriate biospecimen annotation.

> "Standardizations of sample quality, form, and analysis are an important unmet need and requirement for gaining the full benefit from collected samples. Coupled to this standard

is the provision of annotation describing clinical status and metadata of measurements of clinical phenotype that characterizes the sample. Today we have not yet achieved consensus on how to collect, manage, and build biobank archives in order to reach goals where these efforts are translated into value for the patient."

—Standardization and utilization of biobank resources in clinical protein science with examples of emerging applications

The need for standardized annotations is driven by the expanding scope of biospecimen dependent research. Research studies require not only high-quality samples, but also adequate numbers of samples. Sample cohorts are often defined by very specific characteristics and biobanks need to enable the discovery of relevant specimens via a consistent and searchable set of parameters. Precision medicine requires more and more specific specimen subtypes in order to detect the molecular characteristics of disease at an increasingly personalized level. Getting sufficient sample sizes for this requires pooling specimens either from multiple biobanks or collection of specimen across multiple institution or hospitals into a large central biobank. Unless biorepositories are using a common standard for describing their specimens, it becomes difficult to know what is available without consulting each biorepository separately.

Standards are Interoperability Driven

In recent years federated biobanking has become a viable model to increase accessibility of biospecimens across many localized collections. Federated biobanks operate by decoupling basic biorepository logistics, the LIMS aspect (Laboratory Information Management System), from presenting biospecimen to researchers for sample cohort discovery. Multiple biorepositories share a common online database, aka "Virtual Biobank". 5AM Solutions has developed an open source software product called Biolocator to support virtual biobanks.

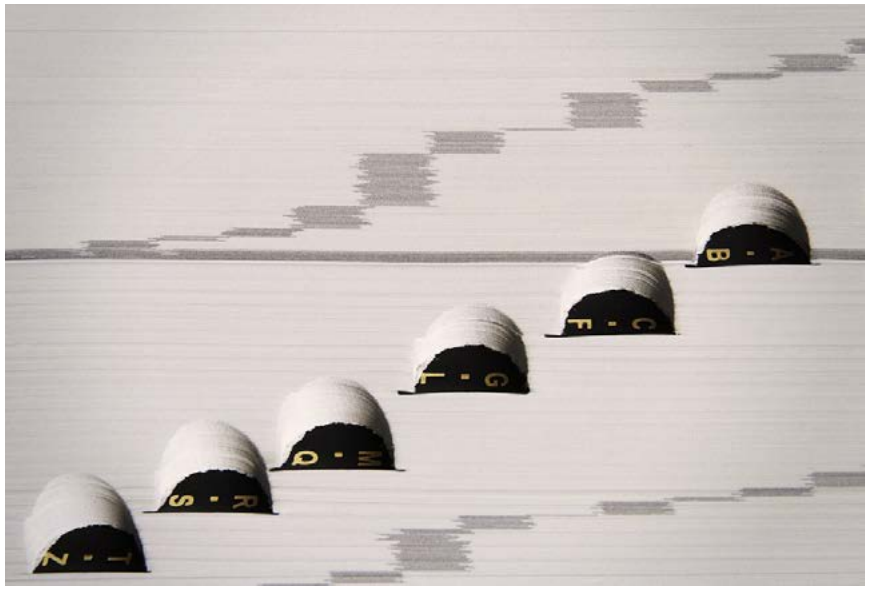
Data standards provide descriptions of the structure of exchanged information such as entity names, data element names, descriptions, definitions and formatting rules. When used for sample annotation, the standards enable interoperability among federated biobanks by guaranteeing that the interacting parties share the same understanding of the shared

biospecimen. They also facilitate data exchange with systems that manage clinical data, for example, EHR (Electronic Health Record) and CDMS (Clinical Data Management System) systems. Data standards are also essential to integrate biorepositories with LIMS solutions and molecular data repositories that manage biospecimen-derived data.

Data Standards are Under-Used In Research Biobanking

Most small scoped biobanks in the research realm use no standards at all when annotating their assets. Even larger biobanking operations map their biospecimen annotations only to few selected controlled vocabularies, which we will discuss later. Key drivers for these selections are interoperability concerns with other departments, institutions, or systems.

The larger the biobanking context the more likely it is that standard vocabularies will be used to map biobanking annotations across biospecimens managed by all participating parties. It is also important to note that current vocabularies used in biobanking are not perfect, but instead provide a starting point for systematic biospecimen annotations. Frequently, local extensions are required to accommodate all shared concepts, as the



standard vocabularies are not always sufficient to accommodate all information contained in modern biospecimen annotations. This is true in particular for many preanalytic parameters, e.g. factors and conditions that influence biospecimen properties and characteristics before analysis.

Data Standards Are Recommended by OBBR and ISBER

Both the Office of Biorepositories and Biospecimen Research (OBBR), and the International Society for Biological and Environmental Repositories

(ISBER) explicitly recommend the use of Common Data Elements (CDEs) and controlled vocabularies for biospecimen annotation to ensure system interoperability and maximum research reuse.

Data should be electronically convertible into formats that can easily be shared among collaborating institutions, where possible and appropriate. The inventory management system should enforce all data integrity, security and audit trail requirements for external access. To achieve interoperability, inventory management systems should do the following:

- Have a public documented Application Programming Interface (API) to enable other systems to integrate with it.
- *Use common public vocabularies for relevant data points (e.g., SNOMED, ICD9-CM, ICD10, ICDO).*

—2012 Best Practices for Repositories (PDF)

- Biospecimen resources should employ a uniform, nonredundant vocabulary (e.g., Cancer Biomedical Informatics Grid [caBIG®] common data elements [CDEs]) for clinical data.

—NCI Best Practices for Biospecimen Resources

While the ISBER guidelines are a good start they lack detail and specificity to truly enable biobanking interoperability. The OBBR recommendation is more precise but references the caBIG(R) standard that, for various reasons, failed to gain widespread adoption in production biobanking solutions. Among them were tight coupling to the caTissue application, lack of sufficient funding to implement (and improve) the standard in the existing biobanking systems.

Previously, we discussed the mandate for using data standards in biobanking. Here we will show how biomedical controlled vocabularies, a kind of consensus data standard, are used to improve data quality and interoperability.Data interoperability standards, eg HL7 and the the ISO 11179 Common Data Element (CDE) standard make use of controlled vocabularies.

Controlled vocabularies, sometimes simply expressed as “vocabularies”, are tools used to standardize information for purposes of capturing, storing, exchanging, searching, and analyzing data. A controlled vocabulary is a restricted list of words or terms used for labeling, indexing or categorizing. It is controlled because only terms from the list may be used for the subject area covered by the controlled vocabulary.

Exploring Controlled Vocabularies

Some of the most well known standard vocabularies created for healthcare are the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), the Logical Observation Identifiers Names and Codes (LOINC®), and the Unified Medical Language System (UMLS®). We will discuss each standard and what they are used for.

SNOMED-CT

The Systematized Nomenclature of Medicine Clinical Terms, or SNOMED-CT, is a general-purpose vocabulary for the medical domain and claims to be "the most comprehensive, multilingual clinical healthcare terminology in the world". It contains concepts in multiple languages about the entire medical domain and provides a hierarchy of those concepts. It also provides multiple terms per concept, allowing lookup of terms with greater or lesser formality. However, the hierarchy is not as rigorous as it can be. For instance, in the top level classes, "organism" is a separate top-level class from "physical object", so this hierarchy should be taken with a grain of salt, rather than used as a firm class hierarchy. SNOMED-CT is therefore useful as a way to align words (terms) from multiple languages to concepts (language-independent, or translatable ideas), which can be used to standardize the annotations or data values of data models like HL7.

LOINC®

The next standard, Logical Observation Identifiers Names and Codes (LOINC®), is more specialized, as it aims to be a "universal code system for identifying laboratory and clinical observations" rather than being a comprehensive vocabulary for the health domain. LOINC is a taxonomy like SNOMED-CT, but is more organized and has nearly-sensible top-level concepts, possibly benefiting from it's narrower scope.

While we haven't taken a rigorous review of the hierarchy, it seems that it is more likely to be usable as a basis for a class hierarchy. However, the top-level concepts seem to be categories of classes, rather than themselves being classes. This is sometimes appropriate for conceptual hierarchies, but bad modeling in class hierarchies. LOINC is also used in HL7 and other data standards.

UMLS®

The Unified Medical Language System (UMLS®) is a "meta vocabulary" published by the National Library of Medicine, providing mappings between SNOMED-CT, LOINC, and many other vocabularies. The UMLS team describes it as "a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems." It is not itself a controlled vocabulary, but is a useful means to find vocabularies appropriate vocabularies for particular applications, and to translate concepts from one controlled vocabulary to others.

Other Vocabularies

There are many other controlled vocabularies used for interoperability standards, including the National Cancer Institute Thesaurus (NCI thesaurus), used in cancer research, the International Classification of Disease (ICD), used to describe diseases, and the Common Procedure Terminology (CPT), which is used to describe clinical procedures.

These vocabularies, and many more, including many ontologies, have been documented and at the National Center for Biomedical Ontology's (NCBO) Bioportal, which also provides tools and APIs for suggesting vocabularies and individual concepts to use as well as tools for authoring new vocabularies and ontologies.

Data Standards Initiatives for Biobanking

Last week we discussed **three controlled vocabularies** that are used in biobanking. This week, we will discuss some data standards that use those vocabularies to make shared data models that can be used to shape and share biobanking data.

CAP Standard For Clinical Pathology

Biospecimen banked by clinical pathology are managed in systems ancillary to electronic medical records and are therefore not unsurprisingly managed by document-based solutions that generate and store pathology reports.

"Content and structure of clinical pathology reports are somewhat standardized by templates and guidelines published by the College of American Pathologists (CAP). For example in 2013 CAP produced 46 cancer protocols as a resource to pathologists to aid in effectively reporting surgical pathology findings necessary to provide quality patient care."

—**The 2009 version of the cancer protocols of the College of American Pathologists: a continuing journey from "Guidelines for Pathologists" to "Standards for Multidisciplinary Comprehensive Cancer Care"**

"Similar to electronic medical records these pathology reports have to be processed by either sophisticated natural language processing (NLP) and or human curators before they can be used for structured biospecimen annotation"

Though just like in their EHR counterparts the structured information in these systems is limited to administrative and billing concerns while pathology and clinical data are part of the narrative often with little use of consistent terminology from CVs. Similar to electronic medical records these pathology reports have to be processed by either sophisticated natural language processing (NLP) and or human curators before they can be used for structured biospecimen annotation. This also apparent in the cancer protocols cited above. SNOMED CT was originally developed by CAP and is used in the protocol vocabularies. However no concept codes are referenced in the PDF or Microsoft Word documents destined solely for human consumption.

NCI Data Standards Initiatives

The National Cancer Institute (NCI) funded two efforts related to facilitate data standards in Biobanking, EDRN and caBIG®. The Early Detection Research Network (EDRN) was initiated in 1998 to improve methods for detecting the signatures of cancer cells. The cancer Biomedical Informatics Grid (caBIG®) intended to link researchers, physicians, and patients throughout the cancer community and was introduced in 2004.

caBIG® ended in 2011 due to an unreasonable focus on the development of ”overly complex and ambitious software enterprise of NCI-branded tools…”, that had only “…limited traction in the cancer community”, as described by the caBIG Board of Scientific Advisors Ad Hoc Working Group.

It is of relevance here that caBIG® also developed the Common Biorepository Model (CBM) to reduce the time and effort required by researchers to locate a biobank that has the specimens they need. The goal of the CBM is to selectively share key information to enable a single search across multiple biobanks.

The CBM supports the idea that data should fit a standardized simple domain model as a means to promote sharing. Even though caBIG® has ended the standard has been implemented by a number of commercial vendors and is also used by the NCI Specimen Resource Locator, a database that helps researchers locate human specimens for cancer research. The5AM Biolocator also supports the CBM.

EDRN is focusses on enabling biomarker detection and requires that specimens be collected, processed, and annotated in a standardized manner and that a set of common data elements (CDEs) be collected with each specimen. CDEs were also an

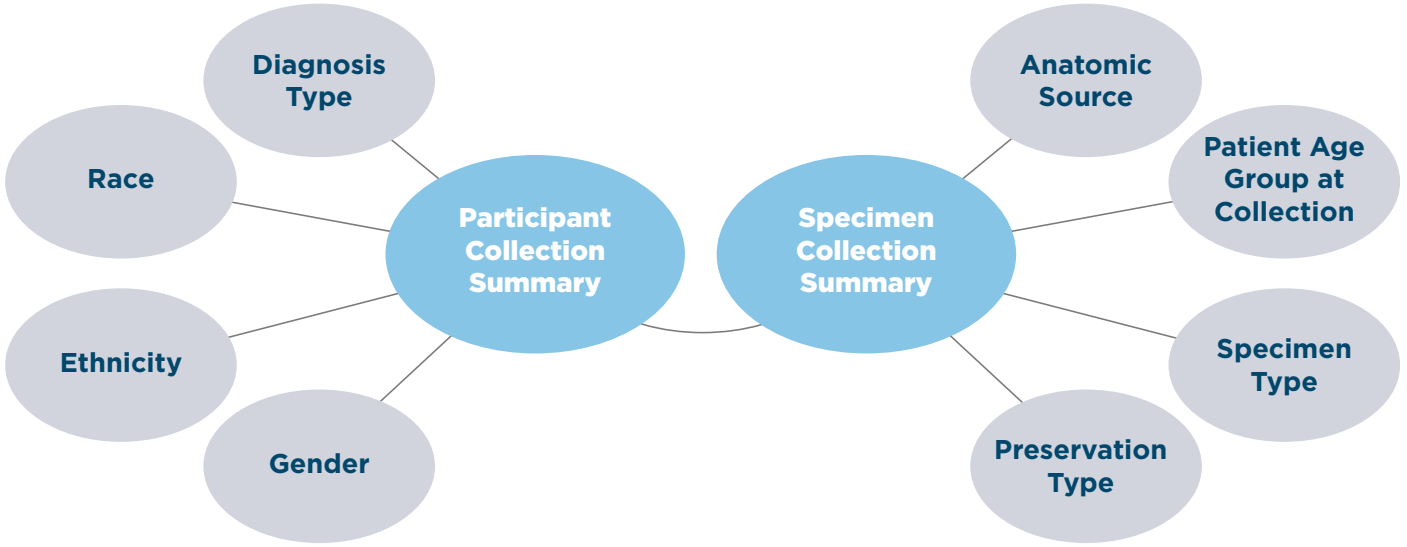


Figure 1: CBM top-level concepts

integral part of the caBIG® interoperability framework and are essentially formalized descriptions of a piece of information. This description contains a name and an exact definition of the specific meaning (semantics—concept mapping) and representation (syntax—data type and format) of this information

BRISQ And SPREC

There have been a number of efforts to develop and introduce specific standards for biobanking. While the standard controlled vocabularies above are more centered in the clinical realm, standards like SPREC and BRISQ concentrate on the formal description of preanalytical parameters. These are parameters like sample collection, processing, and storage conditions that can significantly alter the biospecimens' molecular composition and consistency. Such preanalytical factors can, in turn, influence experimental outcomes and the ability to reproduce scientific results.

Standard PREanalytical Code (SPREC) was developed by the International Society for Biological and Environmental Repositories (ISBER). It iden-

tifies the main preanalytical factors of clinical fluid and solid biospecimens and their simple derivatives in a "specimen barcode". SPREC was introduced in 2010 and is intended to serve as a code that will become recognized internationally within the clinical biobanking sector.

The Biospecimen Reporting for Improved Study Quality (BRISQ) arose from a workshop, Development of Biospecimen Reporting Criteria for Publications, held at the 2009 NCI Biospecimen Research Network Symposium to initiate a discussion on biospecimen reporting recommendations. The list of recommended data elements discussed include general information for consistent documentation of classes of biospecimens and factors that might influence the integrity, quality, and/or molecular composition of biospecimens.

> *"The purpose of reporting these details is to supply others, from researchers to regulatory agencies, with more consistent and standardized information to better evaluate, interpret, compare, and reproduce the experimental results."*

> *"It is hoped that consideration of the BRISQ recommendations will sensitize the biobanking and research communities and their funding agencies to the importance of tracking preanalytical variables, leading to more judicious selection and handling of experimental human specimens and thus improved study quality."*

—**Biospecimen Reporting for Improved Study Quality (BRISQ)**

Conclusion

Despite significant progress in the formulation of biobanking standards, including biospecimen annotations and reporting guidelines, current biospecimen dependent research still suffers from the widespread lack of adoption of these standards by the biobanking community at large.

> *"... lack of international harmonization, uneven adoption, and insufficient oversight of best practices are preventing further improvements in biospecimen quality and coordination among collaborators and biobanking networks."*

—**The Evolution of Biobanking Best Practices**

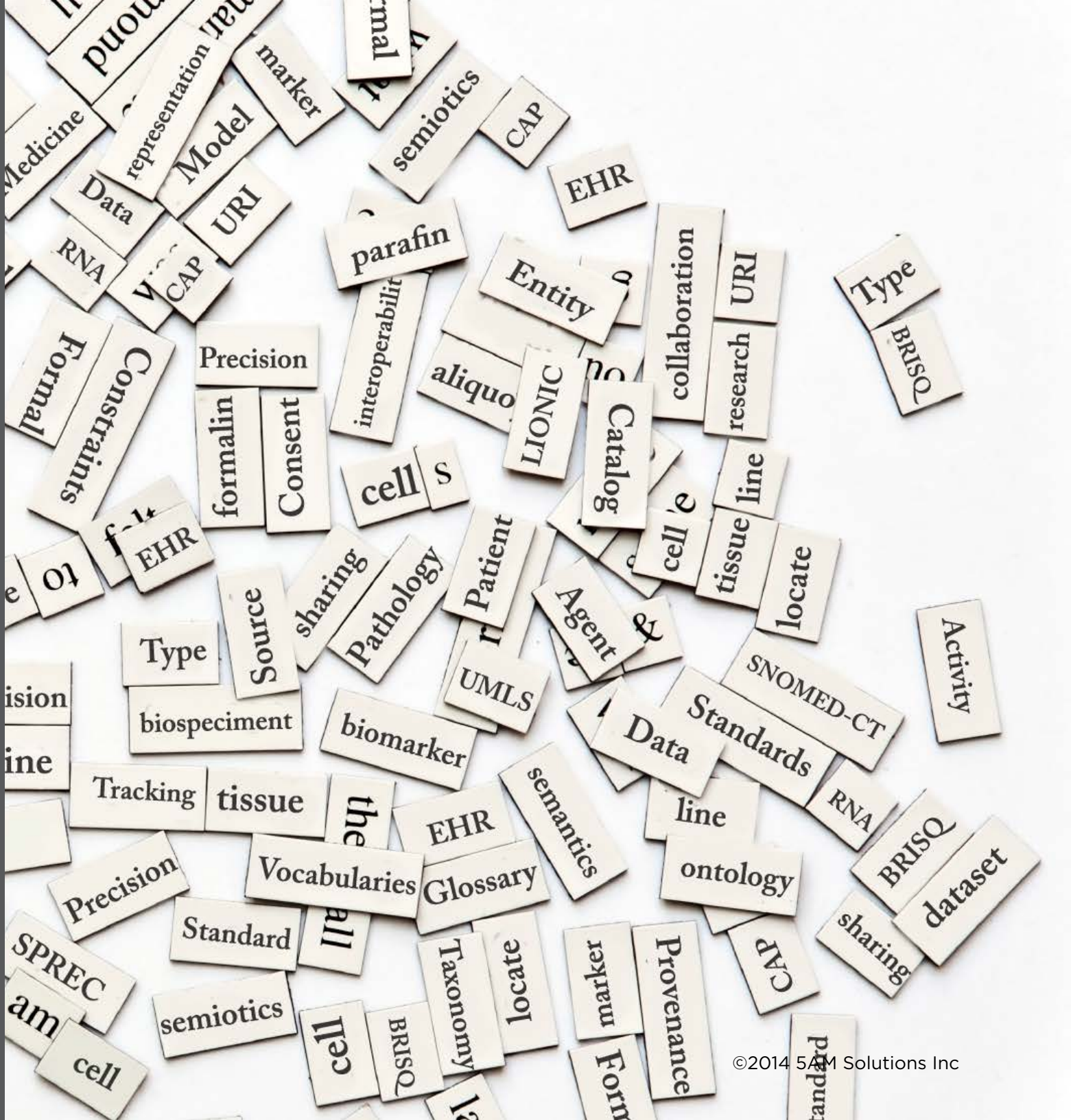
"The lack of data-standards-driven biospecimen annotations essentially secludes millions of valuable samples from an increasingly global biobanking market. According to an August 2012 Infiniti Research report titled "Global Biobanking Market 2011–2015," the biobanking market will increase 30% from 2011 to 2015 to nearly $183 billion."

—Global Biobanking Market 2011-2015

It is time that biobanking administrators, especially in academic medical centers, adopt a more long-term vision when deciding how to manage their biological sample collections. While better sample annotations will require greater investments in infrastructure, logistics and personnel, there will be a significant return on investment in both the economic and scientific sense. Precision medicine, as the name suggests, requires precise information about biospecimen donors, sample collection, processing and storage conditions and systematic capture of sample composition and pathology. Data standards applied to sample annotations at the source of the respective information is a vital component to maximizing the value of biospecimen for biomedical research and translational medicine.

"While better sample annotations will require greater investments in infrastructure, logistics and personnel, there will be a significant return on investment in both the economic and scientific sense."

"It's possible to unlock semantics by mapping your data to semantic web technologies, like OWL and RDF. By doing so, it makes it possible to use your data, along with its built-in semantics, in ways that would otherwise be prohibitively expensive."



Why Semantics Make Biobanks Better

The Semantic Web provides a means to link information on the web to each other and to things in real life in an interoperable way. Internationalized Resource Identifiers, of which URLs are a type, are used to identify nearly everything, and linked data makes it possible to visit those URLs to get more information about the things they represent. This has some very useful applications, especially in biobanking. Semantics was literally made for biomedical research, and here are four ways in which that relationship can help make biobanks better information resources:

Chances Are, You're Already Using Some Semantics

There's a long history between semantics and biomedicine, in fact, some of the oldest controlled vocabularies and ontologies come from biomedicine, and have been driving use cases in semantics since day one. If your biobanking software is using controlled vocabularies like SNOMED-CT, LOINC, ICD, NCI Thesaurus, or FMA, you're already using semantics in your data. It's locked away, and hard to take advantage of, but it is definitely already there.

It's possible to unlock those semantics by mapping your data to semantic web technologies, like OWL and RDF. By doing so, it makes it possible to use your data, along with its built-in semantics, in ways that would otherwise be prohibitively expensive to do (see "Doctors are Patients Too").

Semantics Frees Your Data From Code

This is more general, but there's an axiom in data management: software ages like fish, but data ages like wine. Languages like OWL and RDF let data managers describe the data on its own terms, rather than by how it is used in any one application. Ontologies are, essentially, well-defined data models that provide a universal context for your data. Class membership in a relational database can be expressed as entries in a table, or columns with particular values, or by many other means. It is usually up to software to interpret those expressions.

When data is expressed in RDF and OWL, there is only one way to interpret the class membership property, rdf:type, and that is as membership in a class! This sort of clarity happens because OWL and RDF

use Internationalized Resource Identifiers (IRIs) to describe classes, properties, and entities in your data. When the IRIs used are URLs that return information about those classes, properties, and entities, the data is Linked Data, and becomes self-describing.

A major advantage of this is that many biobanks can use the same ontologies to describe their biospecimens. This provides some additional benefits:

1. It makes it easy to interchange data between systems. Researchers can search multiple biobanks at once without needing to know the details of each system. If the biobank is using an RDF graph database behind the scenes, "easy" becomes trivial.
2. It's extensible both up and down. Implementers do not need to adopt ontologies wholesale, but can simply use the parts that are useful. Similarly, biobank systems can extend their data models using other models they develop themselves or adopt from other organizations.

Provenance is Everywhere—Even in Your Biobank

Provenance is information about how something got to be the way it is. This sort of information is critical to biomedical research, as it can encompass a wide amount of information. Earlier this year, the World Wide Web Consortium, (W3C) released as a recommendation a provenance ontology called **PROV-O** that is intended to be used as a language for expressing provenance on the web. For some basics on using PROV-O, see my **blog post on how newspapers can use it to cite their sources**.

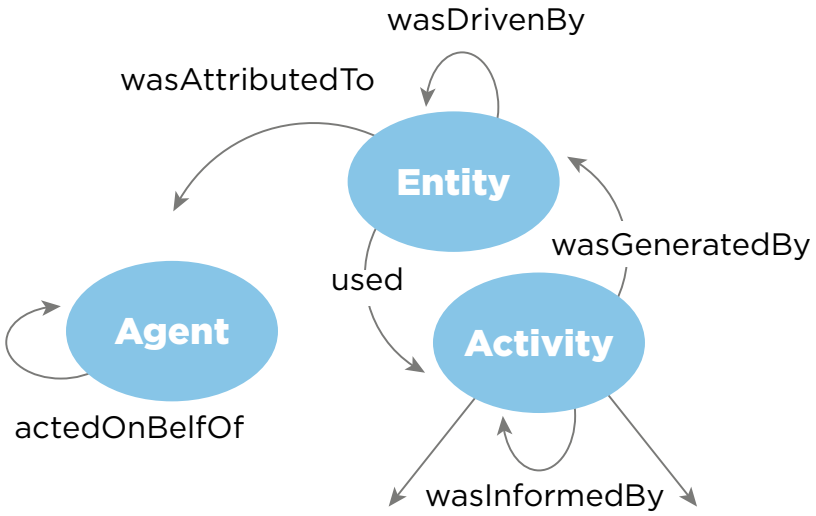
PROV-O is a fantastic example of how data managers and software developers can take advantage of general purpose ontologies. The core of PROV is very simple, and divides the world into entities and activities.Agents can be either entities or activities, but not all entities and activities are agents (this is actually a key feature of OWL, see #5, Doctors are patients too). Entities can be derived from one another, can be attributed to agents, and can be generated by and used in activities.

This core describes a significant amount of work that is done in biobanks, and as people perform work on biospecimens, it is possible to describe that work in terms of PROV-O. Work done in multiple biobanks, when described using PROV-O, can be compared and integrated easily because the mappings are already in place. Since this model has been defined in a global context (anyone can go

and look up the definitions), it is much harder to misinterpret information that uses it.

When PROV-O is combined with other vocabularies that describe specific tissue types and other biomedical concepts, it becomes a ready-to-use biobanking information model that can also be used to describe experimental results, LIMS processing events, and maybe even patient records.

Figure 2: Basic Entities and Relations of PROV-O



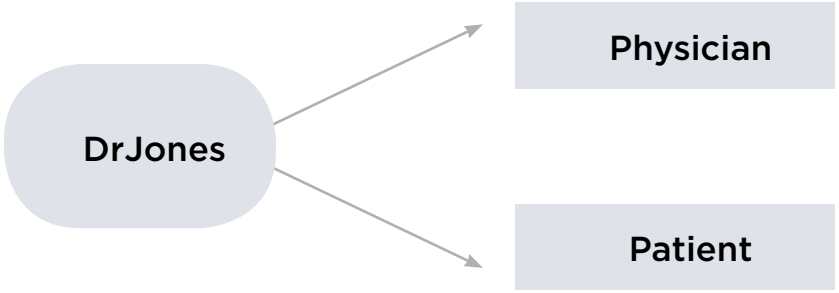
source www.w3c.org

Doctors are Patients Too

In Object Oriented Programming, when an object is made, it gets a particular class. If, for instance, one has a Patient class and a Physician class, it is very difficult to make an instance that is both a Patient and a Physician. In OWL and RDF, it's trivial:

```
DrJones a Physician, Patient.
```

This is because OWL and RDF are expressed as graphs. What we are doing when we say the above is making a graph that looks like this:



That is, the entity DrJones has links that are labeled rdf:type to both the entities Physician and Patient. If there is nothing in the definitions of Physician and Patient that prohibits one from being

the other, this is perfectly fine. The use of graphs is what makes it possible to combine data from multiple sources, as they are simply laid on top of each other. Further, graphs make it easy to talk about other graphs, such as specimen derivation trees, and make it easy to dynamically add new kinds of attributes and annotations on an as-needed basis.

Reasoning over graphs is what makes semantics pretty special—we can create rules that fill out things that are implied by other parts of the graph, for instance, if we say that DrJones has the patient Mr. Brown, we can infer that she is a Physician, simply by the fact that her role in that link is someone who has patients.

Conclusion

Semantics make it easier to understand our data and explain it to others. Embedding those semantics in RDF graph databases makes it easy to share and query that data. When we use ontologies like PROV-O to explain our data, we get a level of free interoperability and mutual understanding that can be very expensive and time consuming to produce through other means. I'll be talking about other interesting uses of OWL and RDF Semantics in the future, so stay tuned.

"PROV-O is a fantastic example of how data managers and software developers can take advantage of general purpose ontologies."

Points Along the Ontology Spectrum

This is the first post in our data modeling series. Today we give a broad perspective for different ways to represent knowledge and data. Some of our posts have talked about ontologies, controlled vocabularies, data models, and other kinds of knowledge representation. All of these share some commonalities, and exist along the Ontology (or sometimes Semantic) Spectrum.

This spectrum was first put in print by Deborah McGuinness in the paper "Ontologies Come of Age" and was developed by her and other panelists at the 1999 AAAI Ontologies Panel. This continuum is defined by its increasing levels of formality, or the amount of additional information that can be inferred from the base knowledge. A quick overview of the differences between the different kinds of knowledge representations along the spectrum should help Batman and Robin resolve their communication issues.

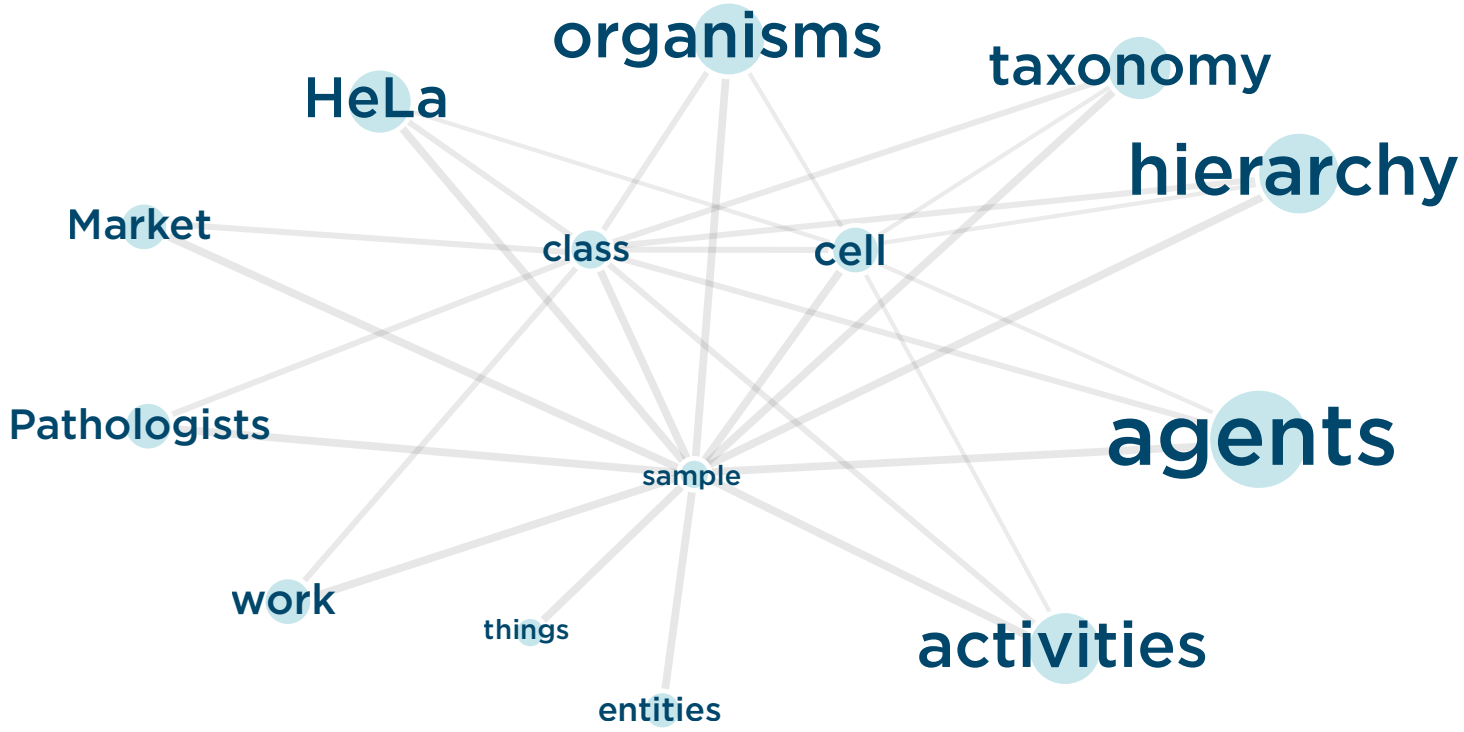
Many of the older ontology spectrum diagrams talk more about the ingredients needed for a position along the ontology spectrum, but here I focus on what kinds of ontologies exist along that spectrum.

Catalog

A catalog, sometimes called a data dictionary, is a basic list of data elements with their permissible values. Usually these elements are locally defined and may or may not include definitions. Data dictionaries are often released with self-contained databases or datasets, andare usually formatted for human consumption in a text, PDF, or HTML document. Tagging systems also fall into this category, since they rely on user-defined tags that can be re-used, but are not often clearly defined by themselves.

Glossary

Glossaries are somewhat more structured, containing identifiers for each term and definitions. Glossaries that aim to be interoperable will use URIs or URLs to identify their terms, but this is unusual. The key difference between a catalog and a glossary is the use of definitions for the terms that it contains.



Controlled Vocabulary

Sometimes called a thesaurus, controlled vocabularies have what is called a weak is-a relationship, sometimes called a broader/narrower relationship. These thesauri relate concepts to each other using predefined relationships, and define a number of terms (labels) that are synonyms for the same concept. The Simple Knowledge Organization System (SKOS) is a well-defined standard that allows the definition of concepts that can have broader, narrower, related, exact match, and close match relationships with each other. We discussed a number of biomedical controlled vocabularies in our Controlled Vocabulary post.

Taxonomy

A taxonomy provides much stronger is-a relationships between entities, and generally talks about actual categories of things. The classic taxonomy, of course, is the biological taxonomy, a hierarchy of classes of organisms, both alive and extinct. In a taxonomy, all members of a class are also members of any superclasses of that class. For instance, all Homo sapiens are mammals. This is expressed in the biological taxonomy by saying that Homo sapiens is-a mammal. In our comic, I'm sure Batman

is correcting Robin because Robin created a class hierarchy that allows for classification, but does not include any additional information, such as what attributes those classes might have.

Data Model

Sometimes called a Frame Data Model or a schema, this is the addition of which properties (attributes and relations) are used by which classes. Classic examples of data models include conventional object models from object-oriented programming languages like Java and C++. In the semantic web, these sorts of models are usually represented using RDF (Resource Description Framework) Schemas, or RDFS, but can sometimes include a subset of the Web Ontology Language (OWL) called RDFS+.

Formal Constraints

Some ontologies take the data model concept further with additional possibilities for logical implications in the form of formal constraints. With formal constraints, classes can be disjoint from each other. Instances, classes, and properties can be declared to be identical to each other, so that if we agree that myNameProperty is identical to yourNameProperty, we can substitute one for the other automatically.

This sort of equivalence, along with the introduction of property restrictions (saying that members of a class must have a value for a property with a particular type, value, or cardinality), means that instances can be classified, or assigned to additional classes based on existing knowledge about them.

Some ontology languages provide a means to write additional logical rules that further extend what can be said in ontologies. Some rule languag-

"If you need to ground your data in existing knowledge about a particular subject, ontologies may provide the greatest benefit. It is important, though, as you look at using different ontologies for your own use that existing ontologies have settled themselves somewhere along this spectrum."

es, like the Semantic Web Rule Language (SWRL), only allow rules that are guaranteed to let the reasoner finish some day (or are decidable, for those of you who took computability). Others, like Common Logic, Datalog, and Prolog, allow for arbitrary rules that are much more free-form, supporting the complete **first-order** (and sometimes other higher-order) logic systems.

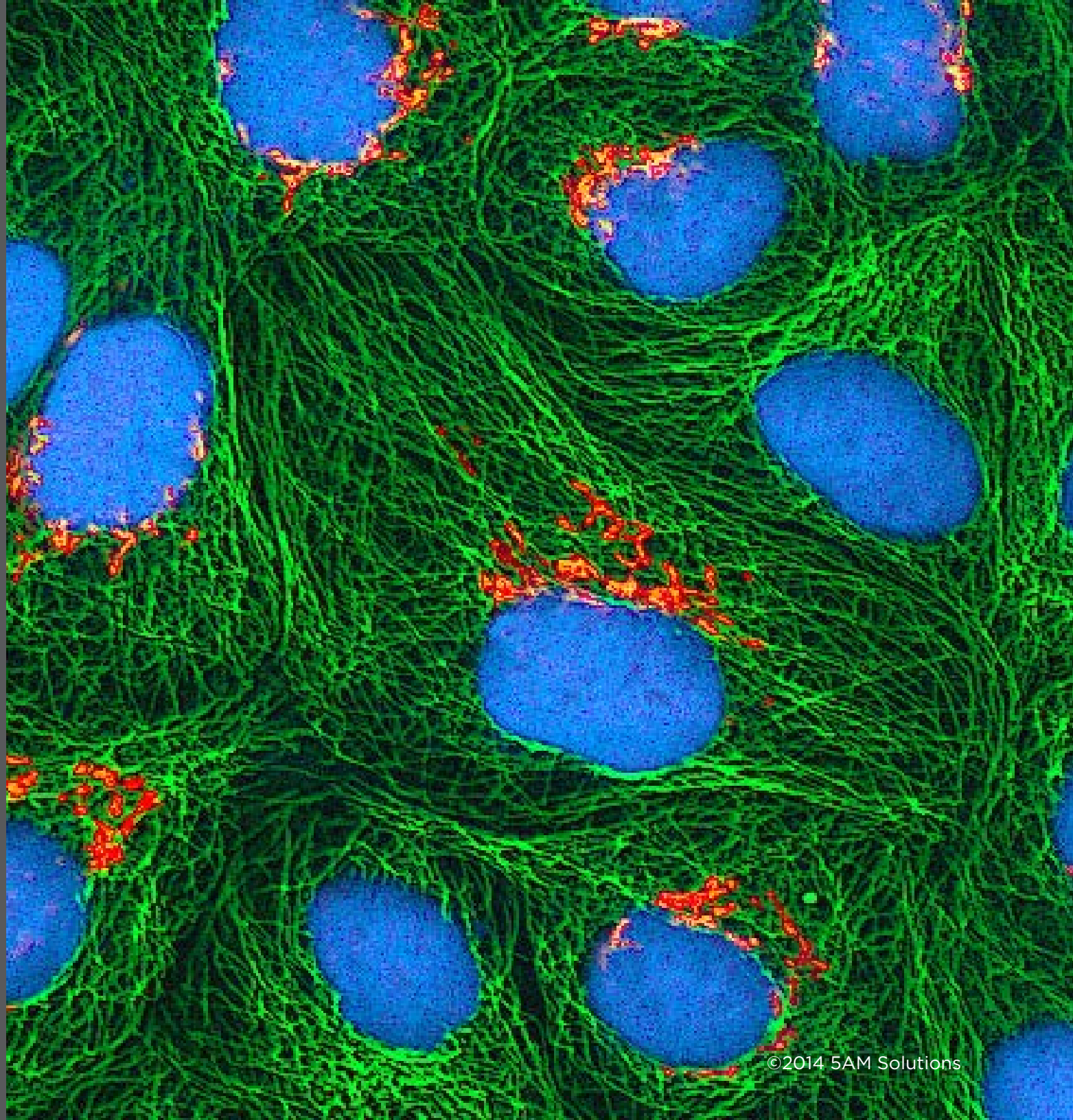
What Should I Use?

This will depend on what you are attempting to do with your knowledge system. If you need to provide subject tags for documents, a controlled vocabulary will work. If you need to create some data that other people can understand, a data model may work well, if it is well documented and easily extensible. If you need to ground your data in existing knowledge about a particular subject, ontologies may provide the greatest benefit.

It is important, though, as you look at using different ontologies for your own use that existing ontologies have settled themselves somewhere along this spectrum.

Their position along this spectrum makes them more or less suited to the use that you have in mind for them—it would not be a good idea to, for instance, build an object model directly from the concepts defined in a controlled vocabulary, or to even create instances of those concepts. Similarly, data models and formally constrained ontologies may not have the vocabulary necessary to perform subject tagging. The value of the ontology is in its use, and pairing the ontology to an appropriate use in your project will help determine its success.

Multiphoton fluorescence image of cultured HeLa cells with a fluorescent protein targeted to the Golgi apparatus (orange), microtubules (green) and counterstained for DNA (cyan). Nikon RTS2000MP custom laser scanning microscope. (Via Wikipedia)



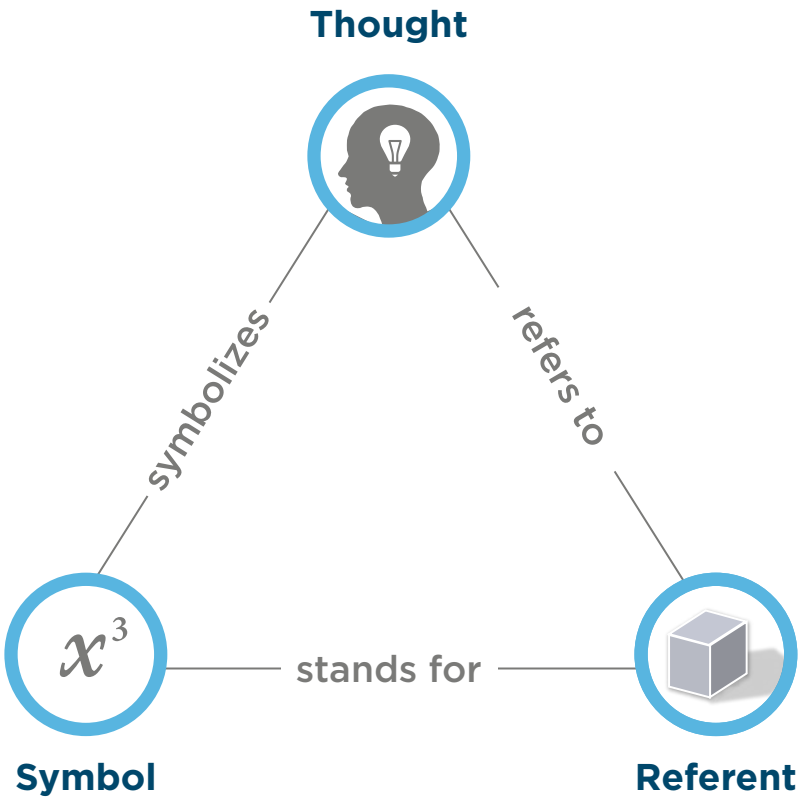
How to Link and Authenticate Cell Lines

Organisms are pretty complicated things, especially multicellular organisms. But what if I told you that, in some ways, single-celled organisms are even more complicated than multicellular organisms? Yet, when we talk about identifying things, this is exactly the case.

A lot of this comes from the way identifiers relate to the things they identify. When we talk about things, we are usually relating three things together: a symbol (a word, identifier, image, number, or something else), a thought (a concept or idea), and the thing itself, called a referent. According to Odgen's Semiotic Triangle, symbols symbolize thoughts, thoughts refer to referents, and through that symbols stand for referents. When dealing with computer systems, it's generally helpful to have a symbol (like an identifier) refer to only one thing, because computers don't do well with ambiguity. Identifying a cell line is complicated because, well, what, exactly are we identifying?

A cell line is a colony of cells that are immortalized from a piece of tissue that came from a multicellular organism. It reverts part of cellular behavior back to things like bacteria and slime molds, but only part.

HeLa is the most famous cell line, and is the most popular both in knowledge share and in cell count. All that popularity means that there are many thousands of cell colonies that have evolved in different ways over the years. Is this still the same HeLa that was extracted from Henrietta Lacks?



"...we need to tell the difference between one colony of cells that have been treated with a particular compound and one that's been left as a control."

In many situations, we would want to say "yes, HeLa is HeLa". But in just as many situations, it's important to distinguish between the modern HeLa cell line and the one originally created in 1951. Further, we need to identify many different strains of HeLa cells, and when we're managing a laboratory, we need to tell the difference between one colony of cells that have been treated with a particular compound and one that's been left as a control.

The most careful position might just say that each and every cell is an identifiable thing, and they are, but that rarely happens unless a single cell has been isolated from a population. Biologists rarely work on this scale, however. Instead, we need a way of identifying colonies of cells and relate them back to the most general cell line.

I wrote a **couple** of **papers** years back that discusses this problem in more detail. The fallout of this has highlighted the differences in perspective that scientists and philosophers have about how to model information, but an outcome has been the addition of two relations to the World Wide Web Consortium's **Provenance Ontology (PROV-O)**, called **alternateOf and specializationOf**. We can use these relations to link different cell colonies together, and also to talk about more abstract representations of cell colonies, like cells that all share a parent colony as its source, even back to the first cells to be immortalized. There are even methods now of **authenticating cell lines**, so these relationships can even be verified and rediscovered.

Biology is a messy business, but it never helps to ignore the mess if instead you can document it.

Summary

As biorepositories have moved away from isolated resources to become more interconnected, data standards and semantics become more and more important. Different methods of expressing semantics, from data schemas and controlled vocabularies to formal ontologies, each provide different benefits and can be employed in complementary ways.

5AM's Biolocator has been tied in with data standards from day one, and has participated in the Common Biorepository Model effort at the National Cancer Institute. As we improve Biolocator, we will be looking to these data standards and interoperability efforts for both guidance for representing and exchanging data, but also as a means to easily leverage data from other providers.

The biorepository is a key piece of the translational medicine pipeline. In the future, the ability of biorepositories to integrate with electronic healthcare records, laboratory information management systems, and research databases will amplify their value and help to create a coherent data strategy for translational medicine.

About 5AM Solutions

5AM enables breakthroughs in healthcare and life sciences through software solutions. Customers save money; connect with colleagues, data, resources and patients; and get better outcomes—*earlier.*

Introduction

5AM envisions, develops and delivers web applications, mobile apps and analytic/collaborative tools to meet the growing needs of the life science and health care sectors. 5AM applies our capabilities in the intersection of technology, science, and medicine. Our focus is to bridge the divides and remove the silos as the domains converge. We employ top software engineers, subject matter experts, repeatable processes and assets built for the domain to solve complex problems with simple solutions. 5AM wants commercial, government, academic, and nonprofit customers achieve new business workflows, interrogate and integrate data, and communicate in new ways. 5AM employs more than 50 employees across the US and is supported by a team of over 30 subcontractor specialists.Services and Software Solutions

Services and Software Solutions

Custom Software—Intuitive user and application interfaces, mobile solutions, data standardization, and legacy modernization are all part of 5AM's daily work. 5AM's track record is driven by pragmatism, technology expertise, and the continuous improvement of our software development methodology. We use our domain experience to guide our teams and customers, share knowledge, and define where we can bring value now and in the future. 5AM's exclusive focus in the domain allows our engineers to produce solutions that are built smarter, delivered earlier, with lower risk and smaller total cost of ownership.

Data Visualization—5AM's software engineers, bioinformaticians, PhDs and MDs with software appreciation and skills specialize in helping customers make sense of large volumes of complex data. We design assays and pipelines. We create compelling user experiences driven off data processing, mining and normalization. We empower more users to ask and answer questions currently accessible only via non-scalable resources and specialists.

Strategy Consulting—From workflow to data grappling, 5AM's consultants help customers develop strategy and scope ways to improve data, tools or front-line operational issues in areas such as enterprise architecture, consortium development and governance. Our ability to conceive what needs to be done, craft a deliberate roadmap of the current and future states of an enterprise, or prototype potential solutions all contribute to the trust customers place in our team.

Packaged Services—5AM has completed so many successful engagements in this domain that we have developed many repeatable solutions. We have been able to invest in evolving many of these components into packaged services that enable us to deliver value-driven and field-tested solutions quickly and at predictable costs.

Competitive Process

In many fields outside life sciences and healthcare, software has reshaped what's possible, particularly in the last 10 years. During those ten years, the convergence of science and medicine has accelerated. This is a complex endeavor; fast changing and highly impactful. Software can't add to that complexity. For years, platforms, point solutions and custom efforts have targeted this translational medicine field. Frustration, shelf ware, wasted time money and opportunity has been a reoccurring result. Software works or it doesn't. Users get final say. And the best way to achieve success is to get the software in their hands. Quickly. Repeatedly. Responsively.

For 5AM, these fields aren't "sidelines." We have built our business reputation by delivering useful, usable and used solutions in the domain for the domain. That focus and our depth of past performance consistently reward clients with faster results, fewer risks and better outcomes. We are familiar with the kinds of user interfaces that will be best-accepted by end users, the challenges of interoperability and the precision reporting required to satisfy regulatory requirements and move faster -- with fewer risks. By collaborating with our customers, we're able to help them produce breakthroughs by using software in new ways, and why 100% of our clients will provide reference for our work.

Customers

5AM is in business to serve the needs of life sciences and health professionals and those they serve. We are customer advocates, coming into commercial, government, academic and nonprofit environments to enable people to ask better questions, find better answers, improve patient care and save lives. Commercial clients include several large Pharmaceutical companies and small biotechs, Life Technologies, Celera, and Quintiles. Government customers include the National Cancer Institute, National Institute of Neurological Disease and Stroke, the National Institute for Child Health and Development, the Office of the National Coordinator for Health IT, the DoD's Transformative Medicine Initiative and the Biomedical Research Commission of Arizona. Non-profits include Janelia Farm of the Howard Hughes Medical Institute, the American College of Medical Genetics, the Translational Genomics Research Institute and the Multiple Myeloma Research Foundation. In academia, we serve New York University, Oregon Health Sciences, UCSF and the University of Michigan's Health System.

Our People

5AM is all about getting our clients better outcomes—*easier and earlier.* Our exclusive focus on software engineering for the healthcare and life sciences domains has enabled us to grow a team of expert developers, requirements analysts, UI designers, and project managers. We also have PhDs, MDs and bioinformaticians with who use deep domain experience to help our technical experts understand the needs and nuances of clients working in scientific and medical fields. This depth of experience and expertise enables us to understand the interests of the divergent set of stakeholders, balance their requests, and elicit and use software to solve problems and unlock all-new opportunities.

We believe in small, expert teams moving quickly to solve customer problems. We believe live software drives the best value and defines working. Ethics and passion about doing meaningful work is required to get in and stay in. 5AM and our customers, teams and partners are rewarded with excellence and value.

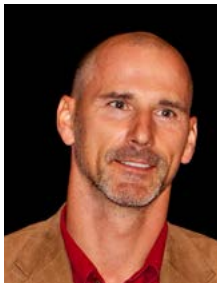
About the Authors

Jim McCusker, Lead Author

Jim McCusker is a Data Scientist at 5AM Solutions and specializes in areas of data provenance and life sciences. He is also a PhD candidate in Computer Science at Rensselaer Polytechnic Institute.

Greg Gurley, Contributing Author

Greg is a senior consultant at 5AM Solutions with over 20 years experience in program and project management and business analysis.



Hannes Niedner, Contributing Author

Hannes is a seasoned biomedical informatics professional who leverages his hands on experience in medicine, bioinformatics and software engineering to connect physicians, researchers and patients with effective IT solutions.

biolocator™

Introducing the world's most innovative system for centralized biospecimen inventory, request management, and order fulfillment. Available in the Amazon cloud. No software to install, fully customizable and ready to run at the click of a button.

Biospecimens are at the core of discovery-oriented scientific research. BiolocatorTM by 5AM Solutions is a centralized, cloud-based biospecimen inventory, request management and order fulfillment system designed to facilitate the finding and sharing of biospecimens located in multiple disparate biobanks to enable scientific and healthcare breakthroughs.

End-to-End Biospecimen Management

Having a way to get the right people the right data is critical. The biolocator provides a framework for moving specimen data from the freezer to the cloud, exposing specified elements for public consumption through a simple de-identified interface. The biolocator manages all information, from biospecimen details to order forms and shipping invoices. All biospecimen data is centrally located to ensure fast and secure search response times, while the biospecimens themselves remain at their original source.

For Researchers and Biobanks

Controlling the use and disbursement of biospecimens is at the heart of the biolocator. Researchers can sign-up and search across multiple biobanks with a click of a button. Once the desired biospecimens are found, researchers can request and have them shipped all from the same interface. For institutions, biolocator provides the ability to scientifically and institutionally review and manage biospecimen requests, configure approved orders, track shipments, and view researcher quality ratings on the biospecimens they have received.

Customized for Your Institution

We've customized solutions for the Arizona Biomedical Research Commission (ABRC) and the National Foundation for Cancer Research (NFCR).

Let the biolocator help you accelerate the pace of research and let 5AM get you there earlier. Contact us at biolocator@5amsolutions.com or (866) 526-6042 today to arrange a demonstration.

Contact 5AM

1700 Rockville Pike, Suite 270
Rockville, MD 20852

(866) 526-6042

Shaun Rabah
Customer Engagement Manager
email: srabah@5amsolutions.com

Subscribe to Our Blog
http://info.5amsolutions.com/