Implementing the National Cancer Institute's (NCI)
Best Practices for Biospecimen Resources:
Bioinformatics and caBIG™ - Next Steps

THE ROAD TO caBIG™ COMPATIBILITY

This document is designed for biospecimen resources that have been introduced to the caBIG™ initiative and are evaluating options and next steps for adopting caBIG™, or becoming caBIG™ compatible. Because different biospecimen resource leaders begin this process with different backgrounds, different sections of this document may or may not be needed to help you decide how to move forward. There are two key sections:

Setting the Stage

Start here if you are still learning about caBIG™ or have hesitation about the benefits of becoming caBIG™ compatible.

- Core Concepts
- How does caBIG™ Compatibility Actually Work?
- Frequently Asked Questions about Bioinformatics and caBIG™

Turning to Solutions

Start here if you want to understand details about the options available to you, their benefits and requirements.

- Software Tools Available from caBIG™
- Overview of Alternatives – Options for Biospecimen Resources
- Skills, Technology, and Resources Required

Resources for Next Steps

The caBIG™ Initiative offers a range of resources for those ready to pursue caBIG™ compatibility or tools adoption.

- **caBIG™ Tools Inventory** (https://cabig.nci.nih.gov/tools) – This webpage lists all the tools currently available through caBIG™ including the caBIG™ biospecimens management tool caTissue Core (https://cabig.nci.nih.gov/tools/catissuecore). Tools pages include documentation and links to download files.

- **caBIG™ Tissue Banks and Pathology Tools (TBPT) Workspace** (https://cabig.nci.nih.gov/workspaces/TBPT) – Provides contact information and resources related to the working group responsible for caBIG™ biospecimen tools.

- **caBIG™ Learning Management System** (http://ncicbtraining.nci.nih.gov/) – Lists the training needed to implement caBIG™ compatibility, including access to the caCORE curriculum, which lies at the heart of caBIG™ compatibility.

SETTING THE STAGE

Core Concepts

There are a number of central concepts in both understanding and achieving caBIG™ compatibility. They are introduced here as a starting point.

Interoperability

- The ability of two or more systems or tools to exchange and use data, or to access the services provided by each tool. There are two equally important requirements for interoperability. First, systems must be able to exchange data. Second, the systems must "understand" the data exchanged. This requires both shared vocabularies, and defined Common Data Elements (CDEs).

Vocabularies

- Vocabularies are an agreed-upon set of standard terminology, available through a common centralized collection. Standard vocabularies are important to any application involving electronic data sharing. These vocabularies provide a structure and a 'same-ness' to the data being collected. For its vocabularies, caBIG™ uses cancer-specific terms in the Unified Medical Language System, captured in the NCI Thesaurus. This is an expanding, dynamic repository of over 60,000 terms; it can be found at http://NCIterms.nci.nih.gov. Each term includes an identification number, definition, source, synonyms, and how the term fits within a larger hierarchy of terms (e.g., how term fits with or encompasses other core concepts). The NCI thesaurus can be searched for both oncology and biospecimen-specific terms.
- Shared vocabularies are a required baseline to match data points in different systems – If the terms in my system match yours, we have shared meaning, so we can link each other's data. The research community grows as a result of this common language. With shared meaning, or shared "semantics," anyone can add a tool and/or data that others can immediately understand.

Common Data Elements (CDE)

- Common Data Elements (CDEs) are standard formats and labels that use the shared vocabulary to define and describe data. When these are the same across different data models and tools, data from one repository can be read and understood by another. For example, if "47" identifies a certain biospecimen, we know "47" is a two-digit number, etc. It might be easier to think of CDEs as the fundamental units of data an organization manages. When an organization needs to transfer data to another organization, CDEs are the units that make up the transaction sets. The transaction set has to include both the data (concept) and how it should be represented (value).
- Databases must be based on CDEs to exchange data with others. CDEs are based on shared vocabularies and are described in a centralized resource called a "metadata dictionary."

Programming and Messaging Interfaces	• Computer programs and the people who write them access resources from other programs through programming and messaging interfaces. Each of these interfaces responds to a particular "code" (or syntax) for its communications. Agreement upon standards for these interfaces is necessary to overcome barriers to interoperability. As noted above, interoperability is the ability of systems to access and use (understand) data. The programming and messaging interface addresses the access part of this combination.
caBIG™ Compatibility	• A "caBIG™ compatible" tool is software that is able to interoperate with other tools in the caBIG™ program. There are four specific areas that are assessed to determine compatibility – all relate to interoperability - including requirements in the areas of Interface Integration, Vocabularies, Information Models, and Data Elements.

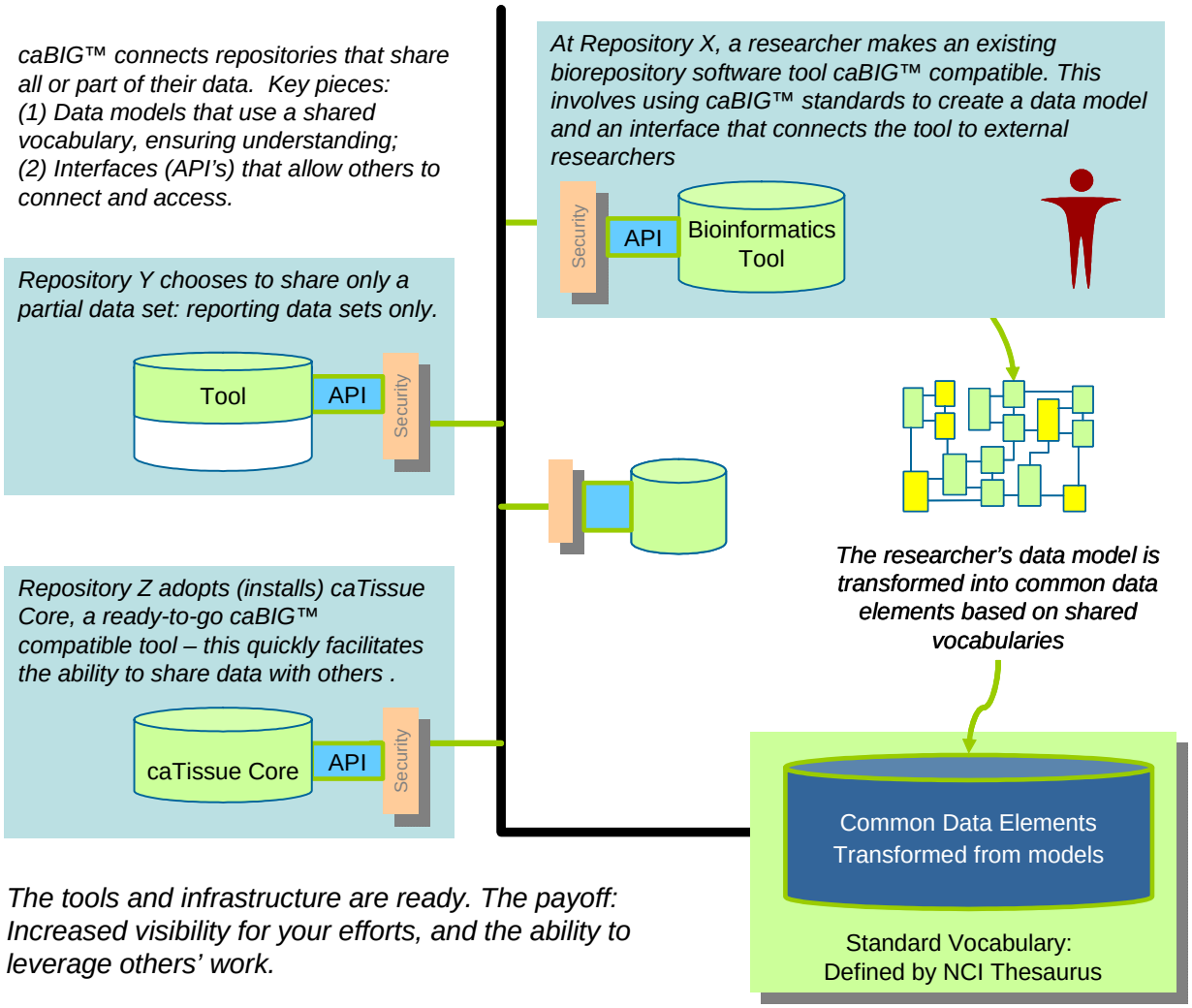
How does caBIG™ Compatibility Work?

The caBIG™ initiative was established to help cancer research organizations collaborate more effectively, by creating a world wide web of cancer research. The benefit of caBIG™ to the biospecimens community is that it lays out a set of ground rules and tools for data sharing between biospecimen resource informatics systems. Tools and data repositories that can share data with others are called "caBIG™ compatible."

The following diagram presents a conceptual view of caBIG™ - all the elements pictured here are ready to go. The centerpiece of the diagram is the caGrid backbone, which when coupled with vocabularies and CDEs allows you to:

- Advertise: Communicate what you have (data and services) and want to share
- Discover: Find out what others have and are willing to share
- Query: Connect with others that are willing to share to get the data you are interested in

caBIG™: An Overview Picture



Frequently Asked Questions about Bioinformatics and caBIG™

The "NCI Best Practices for Biospecimen Resources" has prompted several questions about bioinformatics – especially the implications for smaller labs. Here, we outline our responses.

The Best Practices place strong emphasis on a broad set of clinical data. We don't always have all these data and we don't always need it for our studies, so why is it so essential?

- Expanding the set of data available for a biospecimen beyond the immediate study underway creates the potential for more kinds of useful research. The more contextual data available for a sample, the more potential benefit. For example, while perhaps not essential to your research question, longitudinal data (e.g., survival rates) may be critical to others.
- The more data you have and are willing to share, the more you may be able to access others' data as well. Making your data visible may improve the potential for getting data from others in return – and ultimately can increase trust in the quality of biospecimens from other sources.

The caBIG™ model shows connections between institutions – what about across departments in the same institution?	• Use of caBIG™ compatible software enables integration of systems within the same institution which may currently be unable to exchange data. Such integration can lead to such benefits of improved annotation of biospecimens by linking repositories to Pathology systems and Cancer Registries. This can also distribute the investment costs associated with becoming compatible across multiple departments.
What obligation is there to share my samples once I expose my data? I don't have time to become a tissue shipping department, and I need the samples I have.	• Making your data available does not have to compromise your research autonomy, and you do not have to provide samples upon request. The first priority is your own research project – you will retain control over your samples. • The benefit, however, lies in being able to see what others may be able to provide to help you, and to share what you might be able to provide to support others. This may increase the visibility of your own lab's work, and establish mutually beneficial collaborative relationships with others. • The best practices give you the chance to be recognized for the samples you have – and you have the power to decide how much you want to share these with others.
More generally, why are informatics topics such a significant element of the Biospecimen Best Practices? What's the value to me in implementing these practices?	• Implementing informatics best practices is much like implementing basic security protocols for your freezers – best practices protect both you and your samples. • Biospecimens are valuable resources: both in physical form, and for the related data they provide – data must be maintained as carefully as the associated physical samples. • On a macro-scale, sharing data collaboratively increases public trust in the use of biospecimens and raises confidence that donated tissues are used to greatest benefit. It also communicates that each donation is valued and put to use. • As organizations funded by NCI, following the best practices helps NCI respond to patient concerns and questions related to the use and benefit achieved with the donations received in both financial and human terms.
The focus on interoperability in the best practices seems like a way to force me to invest in new tools. Is this just NCI's way of getting my data?	• We are not asking you to get a new system as a default action. We are asking you to assess whether your existing system aligns with the best practices in terms of functionality, security and other parameters. If they don't, you should identify the gaps and fill them. Adopting existing caBIG™ tools can help. • NCI respects community sensitivities concerning data sharing. Many have expressed concern that data sharing threatens intellectual property considerations, and could hurt a researcher's ability to protect their area of research until they publish their original findings. We recognize that a researcher's top priority is to accomplish study goals, and we are not asking anyone to compromise this commitment. caBIG™ allows you to have control over the data you share. The caBIG™ Data

Sharing and Intellectual Capital (DSIC) workspace continues to take active steps to address issues related to data sharing.

- All this said, there are NIH requirements related to data sharing - and specific expectations related to biospecimens. caBIG™ compatibility helps you meet these expectations without compromising your ability to perform original research using the biospecimens you hold.

TURNING TO SOLUTIONS

Software Tools Available from caBIG™

The caBIG™ Tissue Banks and Pathology Tools (TBPT) Workspace, a special group focused specifically on biospecimen issues, has released two software tools to directly support biospecimen resources. These tools are cost-effective informatics solutions for laboratories and facilities. They are free for download from the caBIG™ website. You may need some technical support in installing them; specific requirements are discussed below.

caTissue Core, Version 1.1	caTissue Core is a ready-to-go informatics tool that covers many functions outlined in the Best Practices, including biospecimen inventory, tracking, quality assurance, and basic annotation. The latest version includes a sample labeling protocol that helps fulfill best practices. caTissue Core is already caBIG™ compatible, and is free. Version 1.1 was released in February 2007.
caTIES	caTIES, a text information extraction system, creates structured data from free-text pathology reports. Data can include histological type, stage and size of tumor, prognostic factors, and molecular markers. The latest release of the tool also includes a de-identification engine that scans free text for Health Insurance Portability and Accountability Act (HIPAA) identifiers, and replaces them with non-identifiable text. Version 2.0 was released in August 2006.

Together, these tools improve the capability for researchers across the country to select and access appropriate samples for their research.

caBIG™ will release the Tissue Banks and Pathology Tools (TBPT) Suite in early 2008, which will bundle all the tools above, and include consent tracking capability in response to user requests.

There are several benefits to adopting caBIG™ tools. First, many biorepository functions encouraged by the best practices are already reflected in caBIG™ tools. Furthermore, caBIG™ tools are designed with security considerations to support patient privacy and data access restrictions, including aspects covered by the “Common Rule” and HIPAA. caTissue Core was built using a Common Security Module – available through caBIG™, which meets best practice goals and intent. Security is also built into caGrid, the caBIG™ architectural backbone. Finally, caBIG™ tools are free – you just need some in-house technical support to get started. Here are some questions that others have asked about caBIG™ tools, and the answers:

What about other vendor tools, like Freezerworks? Are they already caBIG™ compatible?	• Vendors have not been fully assessed for compatibility with caBIG™ - if you have an existing product, we encourage you to ask the vendor about plans for becoming caBIG™ compatible – and to explore whether they will help you attain it.
How stable are the existing caBIG™ tools?	• The software released through caBIG™ is fully tested, and the caBIG™ team continues to respond to needs as they emerge. The latest releases of the caBIG™ tissue banking tools address many user responses to previous releases. For example, a previous release of caTIES required the purchase of a commercial de-identification tool. Now, you can access free tools instead, mitigating the prior concern.
I've heard that caBIG™ is in a "pilot" phase, and not really ready to go. What's available now?	• The central tools of caBIG™ - the vocabularies, the development processes and toolkits, training, and a variety of tools and services on the Grid - are ready today. The extent to which caBIG™ is ready for *you* – in terms of a match to your interests and needs – depends on your requirements. • Like any collaborative or scientific exploration, caBIG™ will never be "finished." It continues to grow as more get involved. • The caBIG™ tools for biospecimens – caTissue Core and caTIES - have been tested, fielded and adopted by multiple organizations. For organizations with advanced informatics capabilities, they can even be extended for your needs.

Overview of Alternatives – Options for Biospecimen Resources

Biospecimen resources represent a variety of operational environments, ranging from very small labs operating with only one or two freezers, to large labs with existing sophisticated informatics systems. Given this diverse environment, caBIG™ offers a number of options for different existing scenarios. This following table outlines alternatives for moving from your current state to caBIG™ compatibility.

Options for Becoming caBIG™ Compatible: Different Solutions for Different Scenarios

Option	Operating Scenario	Recommended Solution	Required Process
1	Your resource has a paper-based system or a homegrown tool that does not meet best practice standards, and which would not be painful to abandon, as long as existing electronic data could be migrated to a new tool.	Adopt caBIG™ software tools – in particular, caTissue Core. This tool meets many of the best practices already, and adopting it automatically makes you caBIG™ compatible.	Adopting caTissue Core will require dependent software, such as JBoss and an underlying database. You may need some technical help to install and maintain caTissue Core .
2	Your resource has an existing basic tool that you feel strongly about keeping. Examples might include an Excel sheet or Access or mySQL database that you don't want to abandon.	You can become caBIG™ compatible by installing caTissue Core, and then mapping your existing tool to it. Connecting to the outside world through caTissue Core automatically makes you caBIG™ compatible.	This solution requires either manually replicating data between the two tools (your existing tool and caTissue Core) – or - building an Application Programming Interface (API) to connect and exchange data between your existing tool and caTissue Core. caBIG™ can provide sample APIs to start this process.
3	Your resource has an existing informatics tool that is more complex than a simple spreadsheet or database. Your tool has separate modules for standard reporting and for data storage (e.g., reported data represents a small extract from a larger database system).	Make the existing tool caBIG™ compatible for your standard reporting structures only. The result: data generated in required reports would be caBIG™ compatible; the underlying data source need not be. If reports are standard, this is a good solution; if your reporting requirements vary, it may be easier to convert entire tool (see Option 4).	Develop and annotate a UML model, convert to CDEs, and create an external API (see Option 4 for more detail). While more intensive than adopting a caBIG™ tool, it is a lighter-weight solution than making your entire tool/database compatible, with the work focused specifically – and solely – on reporting structures and needs.
4	Your resource has an existing complex informatics tool (like Option 3), but your reporting needs vary greatly, and you would just like to have the entire system caBIG™ compatible to allow for maximum flexibility.	This is the highest investment solution, and involves making the full database compatible, by creating an interface that maps the existing tool's data structures to caBIG™ standards.	This represents the most complex solution – depending on the complexity of your tool, it could take weeks to months for an IT team to complete. Basic steps include: building a UML Model, annotating the model using caBIG™ controlled vocabulary (NCI Thesaurus), converting the UML Model to CDEs, and creating an external API using the annotated model.

Skills, Technology and Resources Required

You should not have to hire full-time staff, or invest in an IT lab, to fulfill the bioinformatics related best practices. You may need to "borrow" personnel at your institution with some specialized IT skills to accomplish initial development and perform routine maintenance.

All of the options above require some baseline technical skills, including intermediate programming and database development skills. Someone with intermediate Java and/or .NET development skills should suffice. They also need experience with installing dependent software sets (e.g., Java, JBoss, MySQL), administering systems, configuring line code, and similar tasks of moderate technical complexity. The required technical environment will include application and database servers. These resources should be available at your institution without heavy investment by your lab.

As a baseline, Options 1, 2, or 3 described above shouldn't require more than a few weeks effort depending on your existing tool. Full system conversion (Option 4) is likely to take more investment – on the order of months rather than weeks.

It's difficult to calculate the precise dollar amounts that it takes to accomplish the above options based on different starting points. As more people implement in different environments, we will be able to provide a more accurate assessment, as well as lessons learned and tips for those starting the process.

Tools and training are available from caBIG™ to support the adoption and compatibility process:

- The caCORE Software Development Kit (SDK) and Developer's Guide specifically support data model development. See http://ncicb.nci.nih.gov/NCICB/infrastructure/cacoresdk.
- caBIG™ Boot Camps are held periodically to walk through the steps to compatibility – these are generally 1.5-2 days, and are held at NIH. Watch the caBIG™ website for upcoming events.
- The caCORE training curriculum offers a variety of classes covering different aspects of caBIG™ compatibility. http://ncicb.nci.nih.gov/NCICB/training/cadsr_training/CourseOffering
- NCICB Applications Support can answer tool-specific questions if you encounter questions during caBIG™ tool installation: ncicb@pop.nci.nih.gov.

CONCLUSION

One of the most challenging areas in the field of biorepositories has been the absence of information technology tools and infrastructure to facilitate the appropriate collection, processing, archiving, and dissemination of biospecimens in the research community. Linking researchers, physicians, and patients throughout the cancer community, caBIG™ is the cornerstone of NCI's biomedical informatics ongoing work to transform cancer research into a more collaborative, efficient, and effective endeavor.

For more information:

- caBIG™ Website for Developers and Adopters: https://cabig.nci.nih.gov/
- Public caBIG™ Website for General Audiences: http://cabig.cancer.gov