Paul A. Fearn, MBA
NLM Informatics Research Fellow
University of Washington, Seattle, WA

Informatics Framework for Biospecimen Science

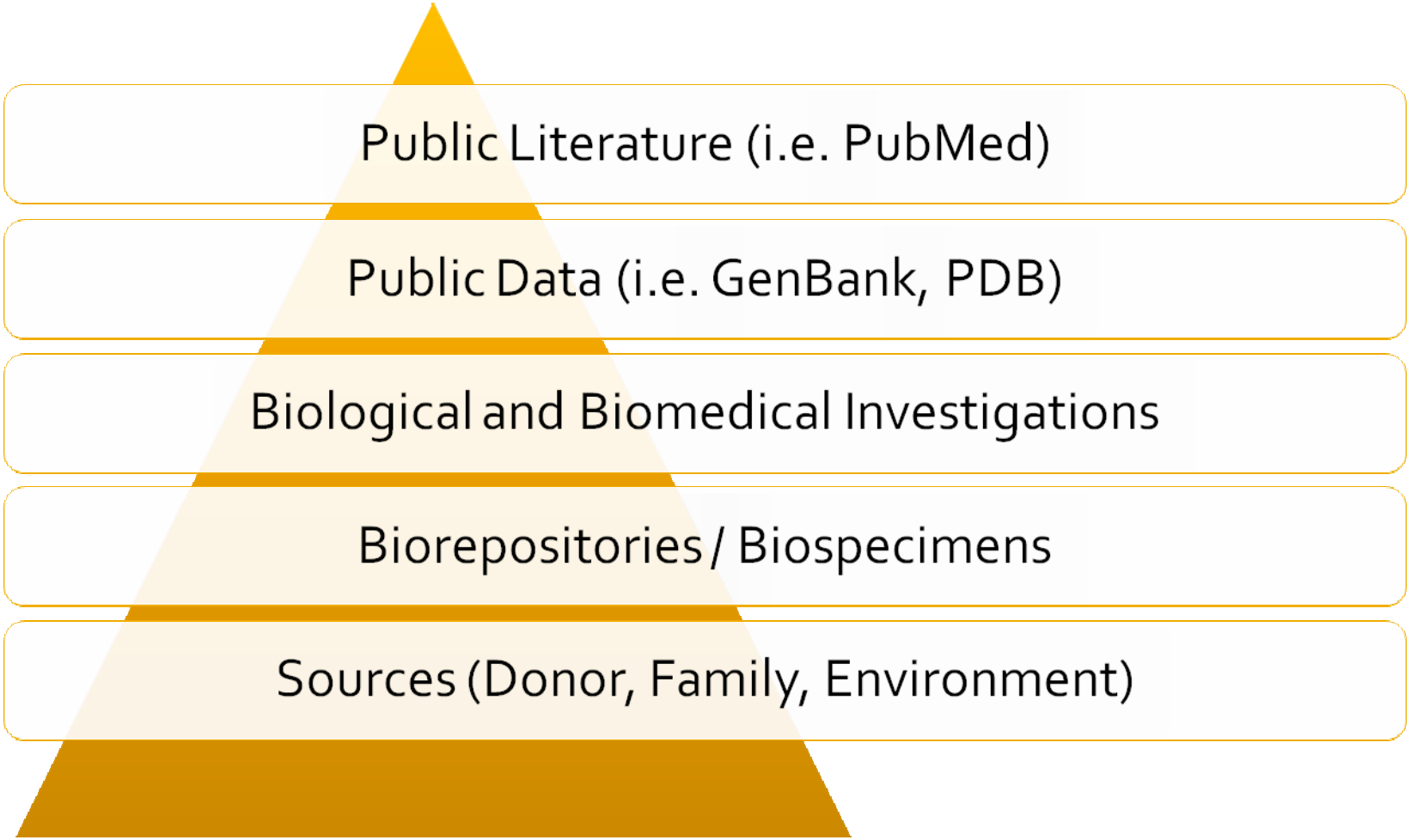
This research is supported in part by the NLM training (NIH NLM #T15 LM07442) and ITHS (NIH NCRR 1 UL1 RR 025014) grants

Goals

- How do we capture, store, retrieve and keep (or link) all necessary data / information for biorepository and biospecimen research?

- What is the scope of informatics for this area?
- What are the boundaries between systems?
- What standards and systems already exist?
- What are the informatics gaps and opportunities?

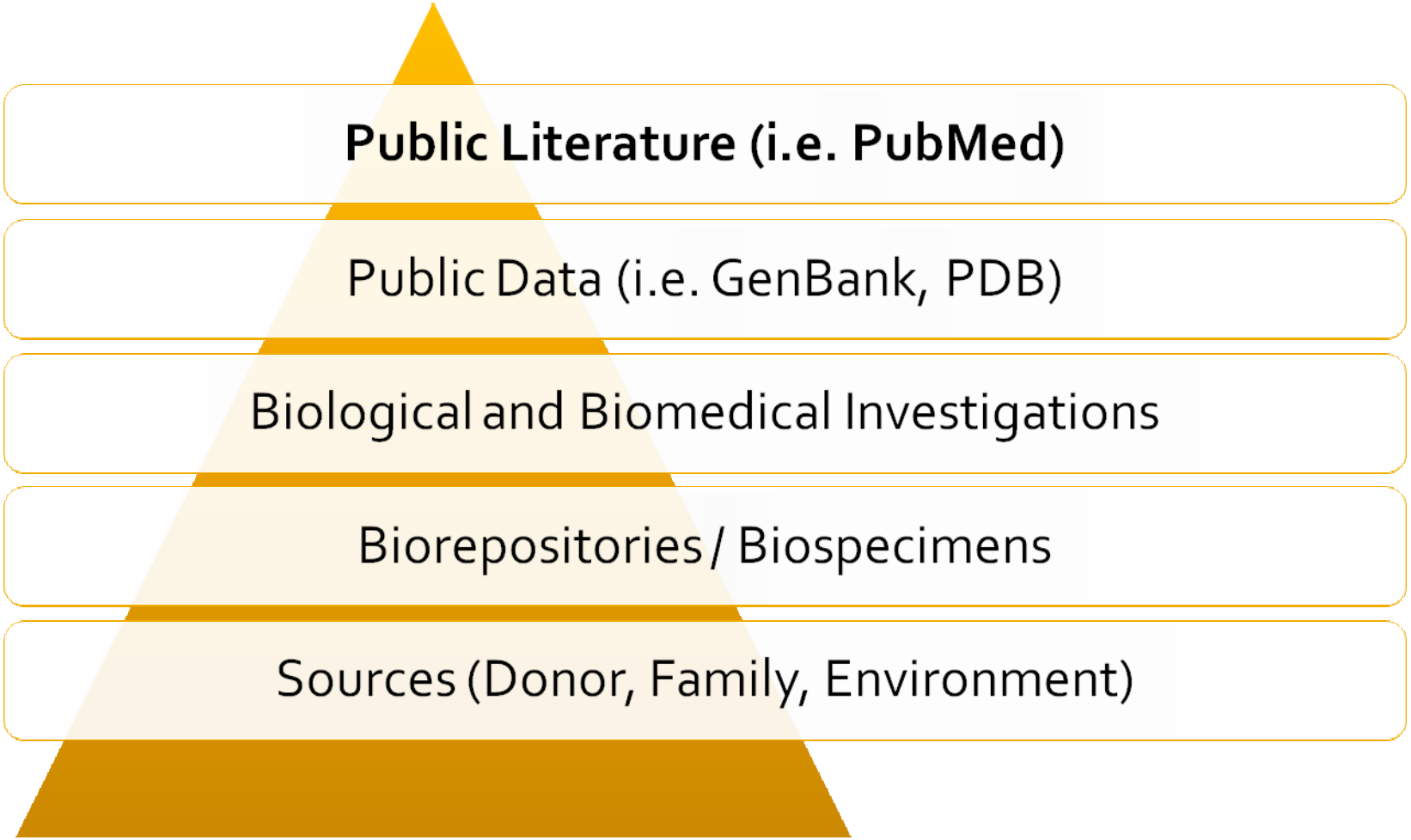
Informatics Framework Components



Methods

- NCI OBBR Biospecimen Research Database
  - Biospecimen science
  - Search terms / indexing
- Literature Review
  - Standards for biological and biomedical investigations
  - Biorepository informatics

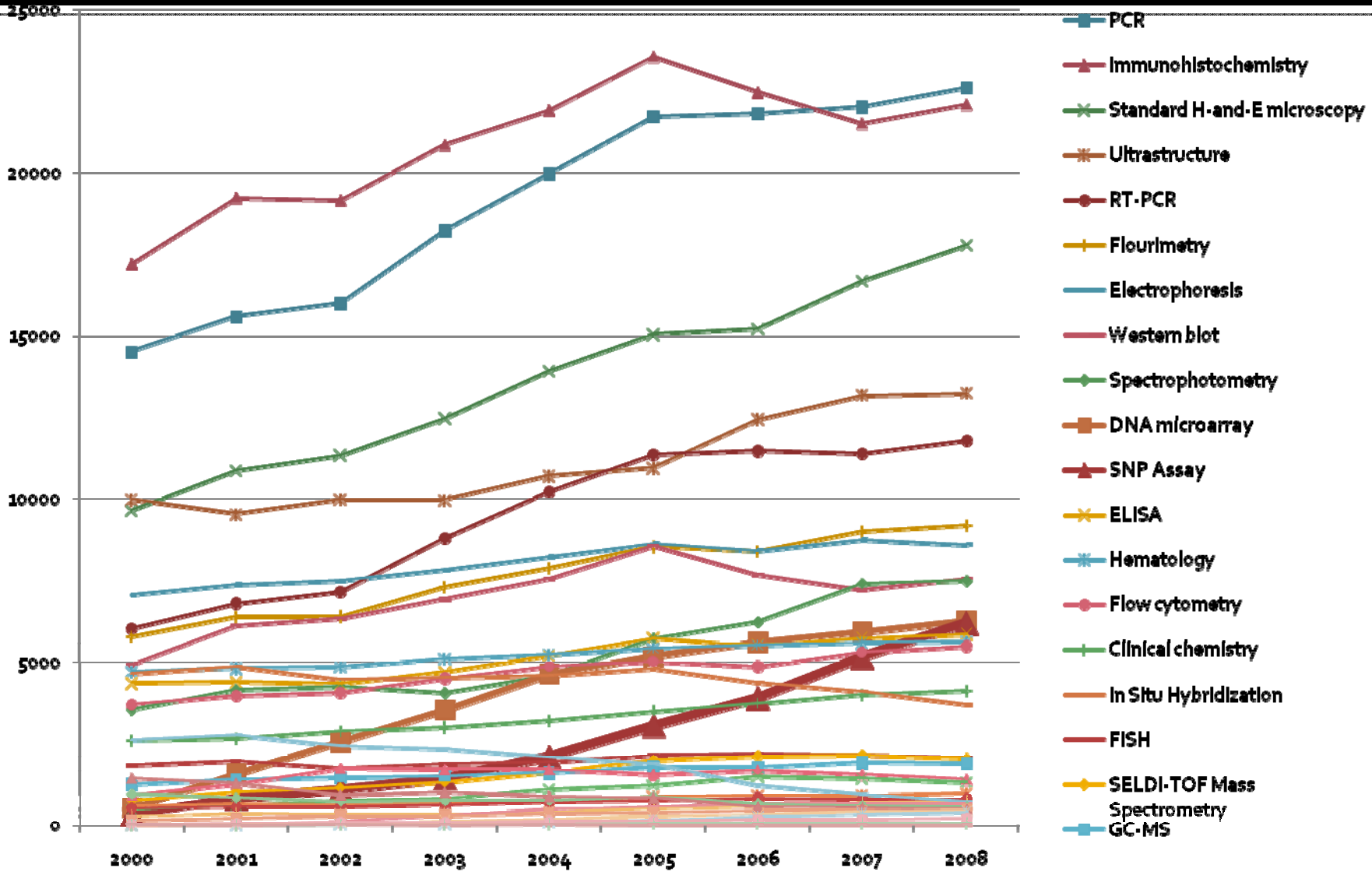
Informatics Framework Components



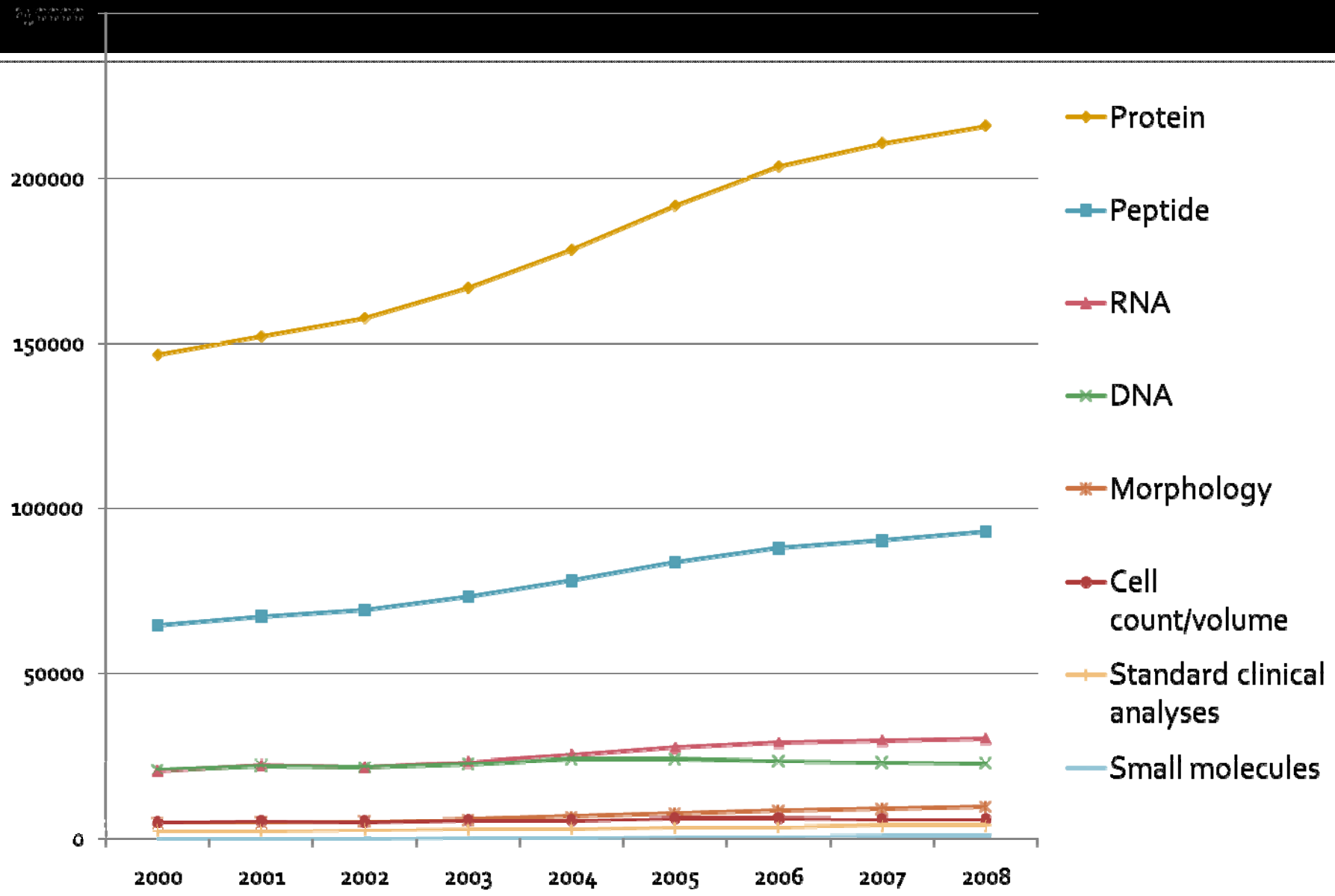
BRD Terms and Public Literature

- https://brd.nci.nih.gov (NCI OBBR web site)
- Search PubMed for biospecimen science
- Extracted and curated information / SOPs
- BRD search criteria / terms
  - **Biospecimen type**: blood, cell, fluid, tissue
  - **Analyte**: DNA, RNA, protein…
  - **Technology platform**: PCR, FISH, CGH..
  - **Location**: blood, serum, plasma, pancreas, bladder
  - **Diagnosis**: melanoma, prostatitis
  - **Preservation type**: formalin, frozen
  - **Experimental factor**: deparaffinization, ischemia time

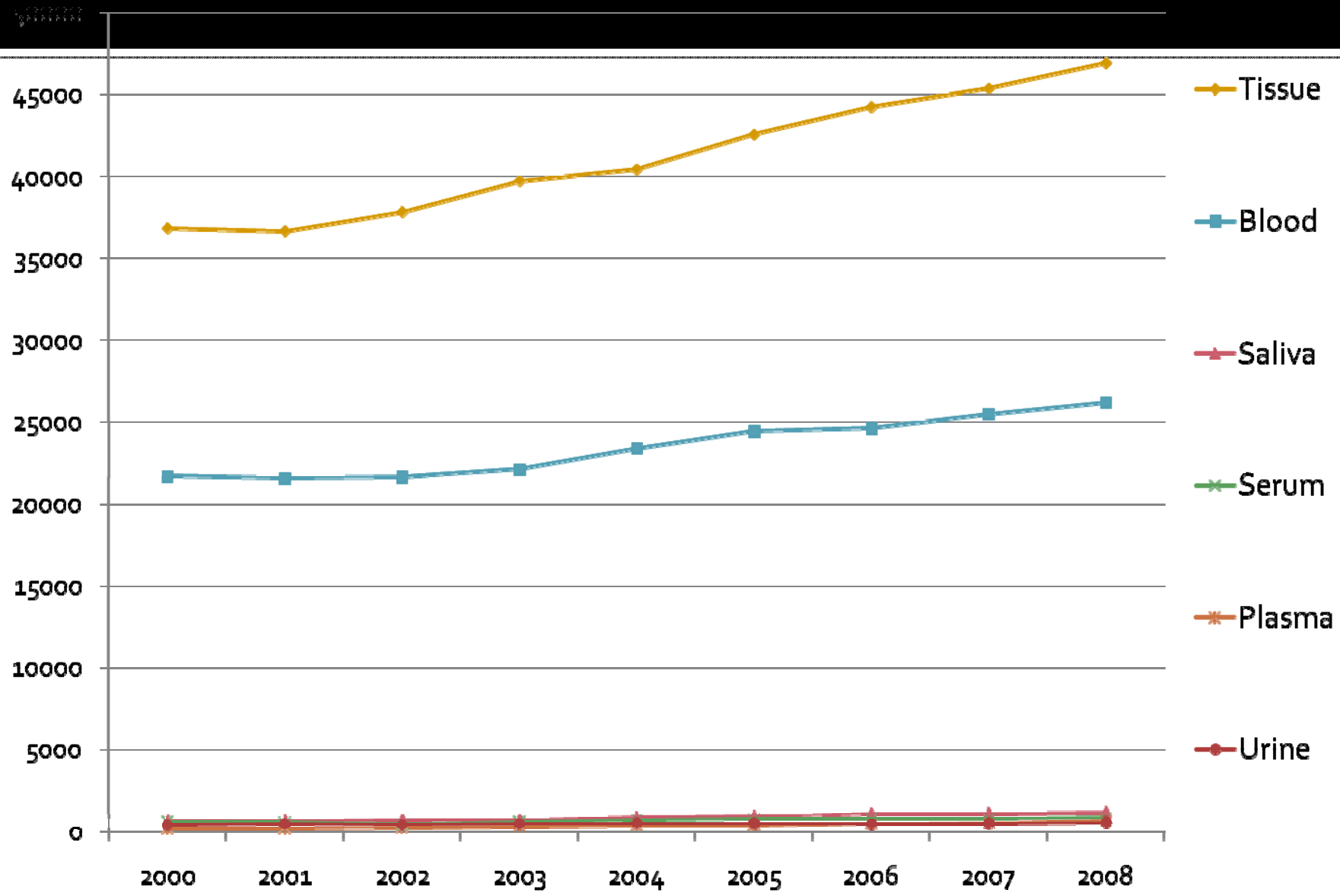
Technology Platform Trends in PubMed



Analyte Trends in PubMed



Biospecimen Type Trends in PubMed

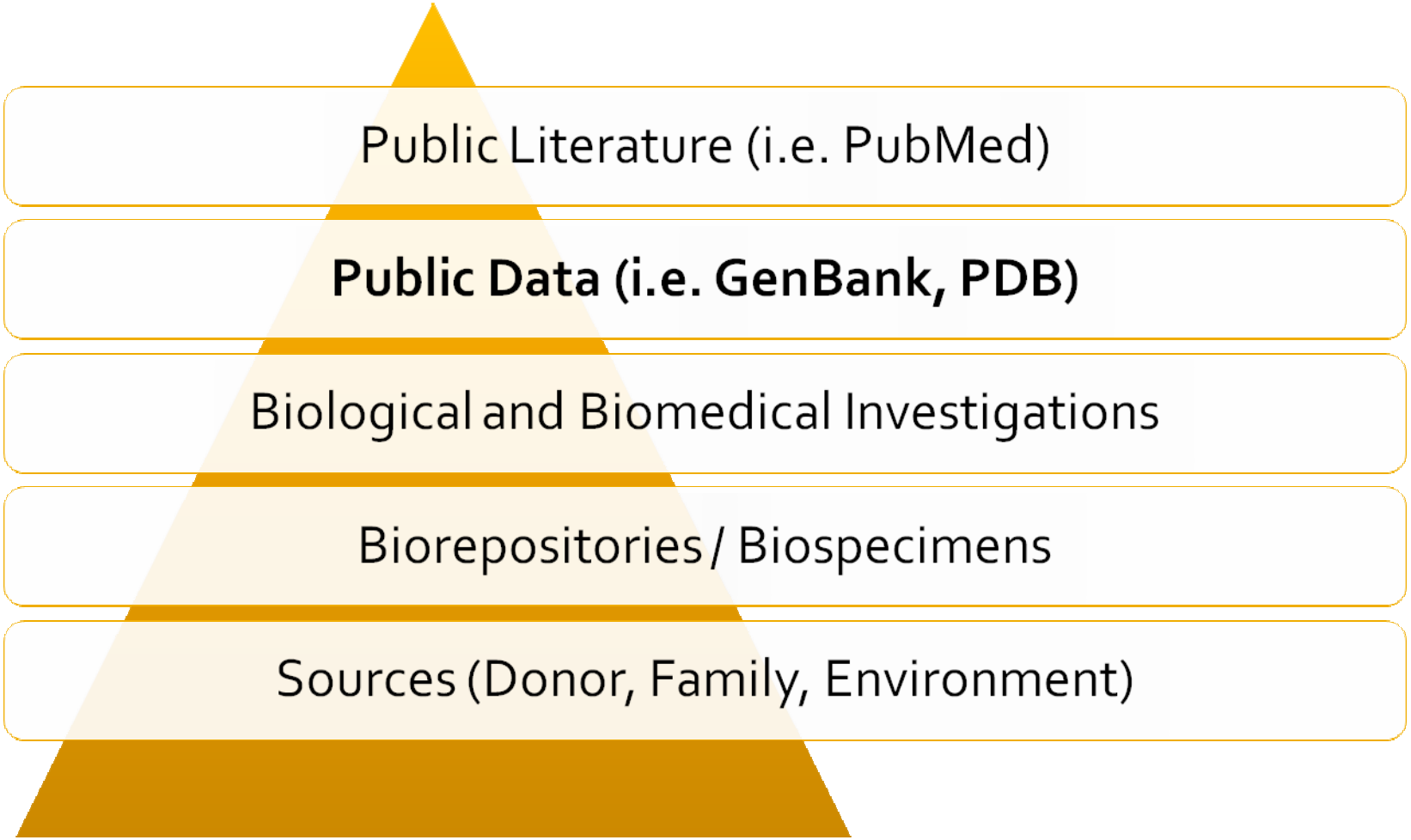


Public Literature Summary

- Tracking expression of samples in literature
- Alignment with BRD of biospecimen science
- Track provenance back to to specimen/source

- Informatics problem: harmonizing and aligning the terms across framework

Informatics Framework Components

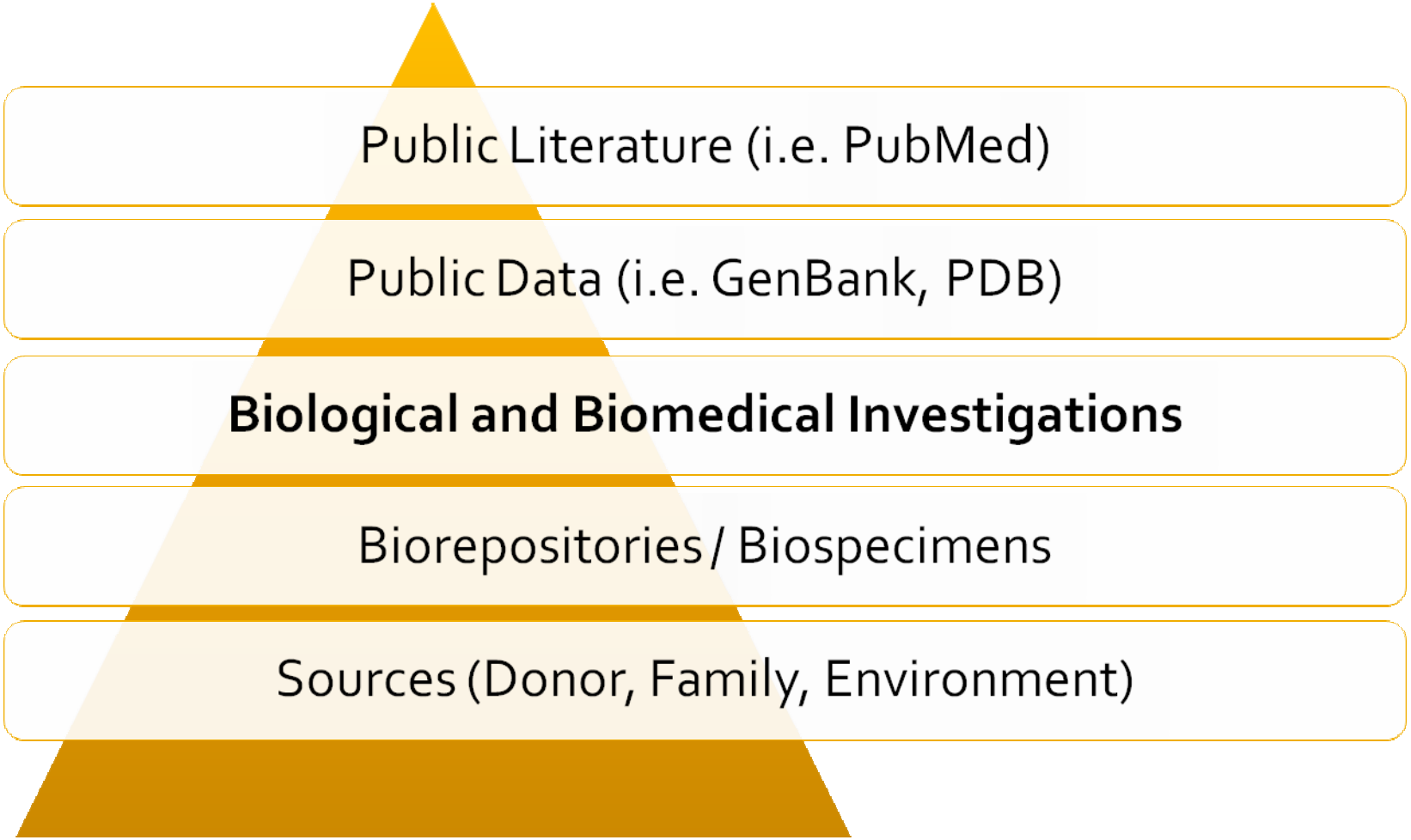


Data from Investigations on Samples

- Variety of public databases (e.g. GenBank, GEO, Peptidome, PDB)

- How is the sample information reported?
- Information to determine provenance?
- Can we determine biospecimen collection, processing, and handling protocols and variations in generation of these data?

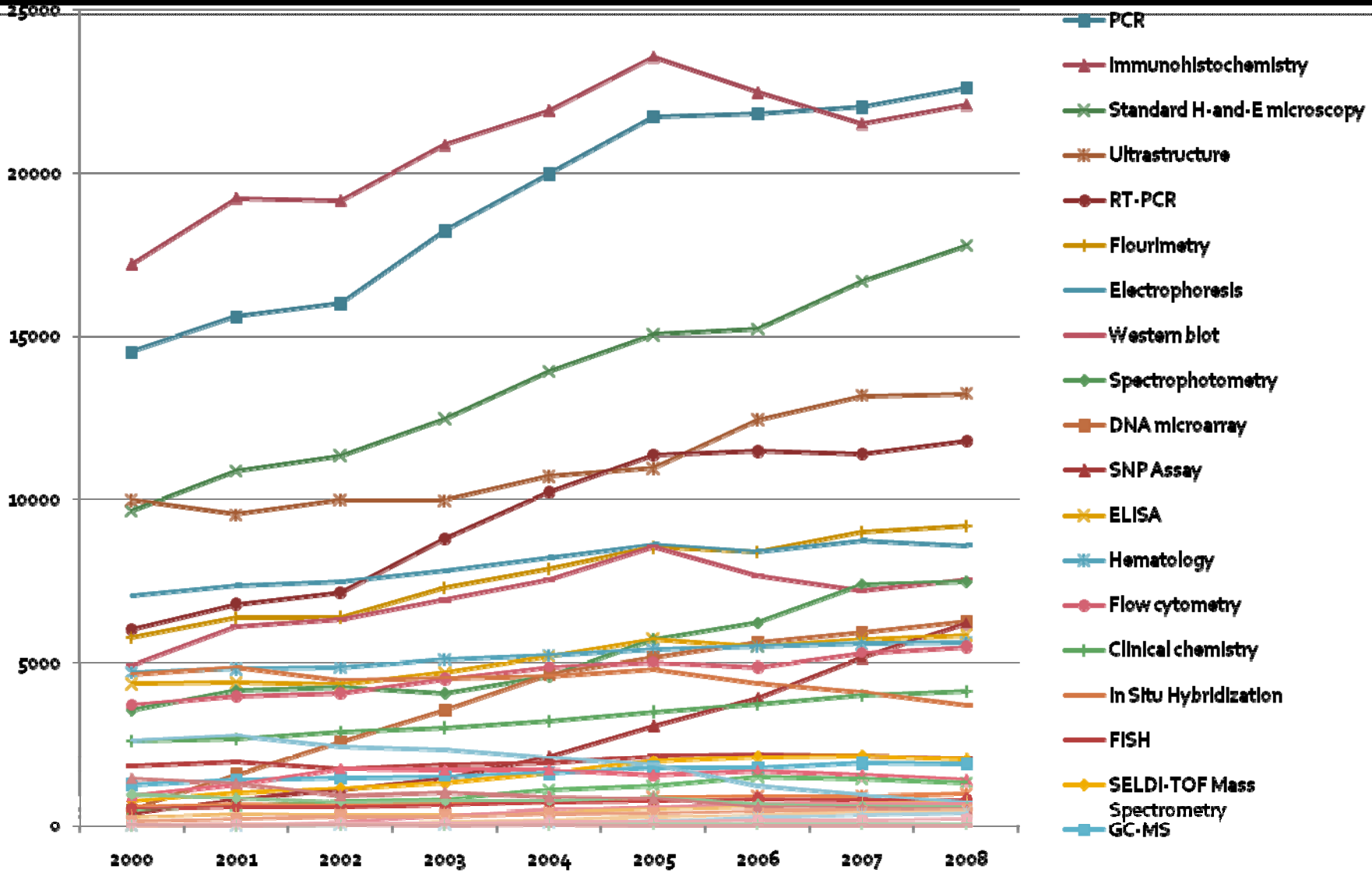
Informatics Framework Components



Investigation Guidelines & Standards

- "Evolution" of standards for biological and biomedical investigation
  - Checklists (MIAME to MIBBI)
  - Format: XML and TAB (MAGE-ML to ISA-TAB)
  - Terminologies / ontologies (MGED-O, NCIt, OBI)
  - UML Data / Object Models: (MAGE-OM to FuGE)
- Coverage of BRD technology platforms?
- Alignment of sample annotation?

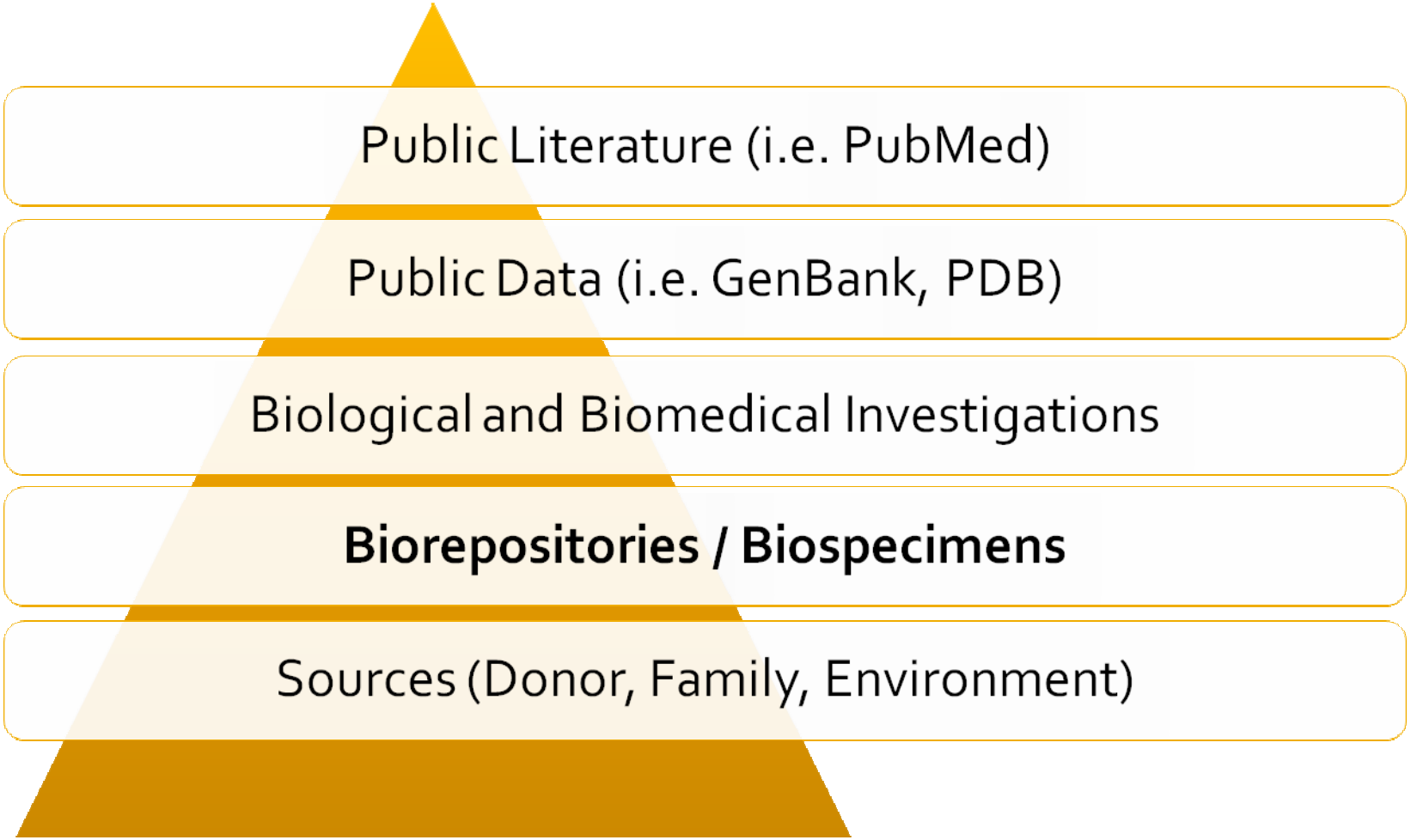
Technology Platforms from BRD



Technology Platforms & Standards

Technology Platform	MIBBI	Data Format	CV / ontology	Model
DNA Sequencing	MIGS, MIAME, MINSEQE	MAGE-ML, MAGE-TAB, ISA-TAB		
FISH	MISFISHIE	ANISEED, COMPARE	OBI	FuGE
PCR	MIAME, MIqPCR (MIQE)	RDML	SO	
DNA Microarray	MIAME	MAGE-ML, MAGE-TAB	MGED Ontology	MAGE-OM
Electrophoresis	MIAPE-GE	GelML, AGML	OBI	FuGE
Immunohistochemistry	MISFISHIE	ISA-TAB	OBI	FuGE
GC-MS	MIAPE-MS	mzML		
Tissue microarray		TMA DES, TMA-TAB	Stanford	TMA-OM
GC-MS	MIAPE-MS	mzML		
Standard H-and-E microscopy	OME			
Flow Cytometry	MIFlowCyt			FuGE

Informatics Framework Components



Informatics Best Practices & Guidelines

- International Society of Biological and Environmental Repositories (ISBER)
  - Best Practices for Repositories, 2008
  - Revised version in progress
  - Human and non-human data elements
- NCI OBBR
  - Best Practices for Biospecimen Repositories, 2007
  - Revised version in progress

http://www.isber.org/ and http://biospecimens.cancer.gov/

Basic Biospecimen Tracking

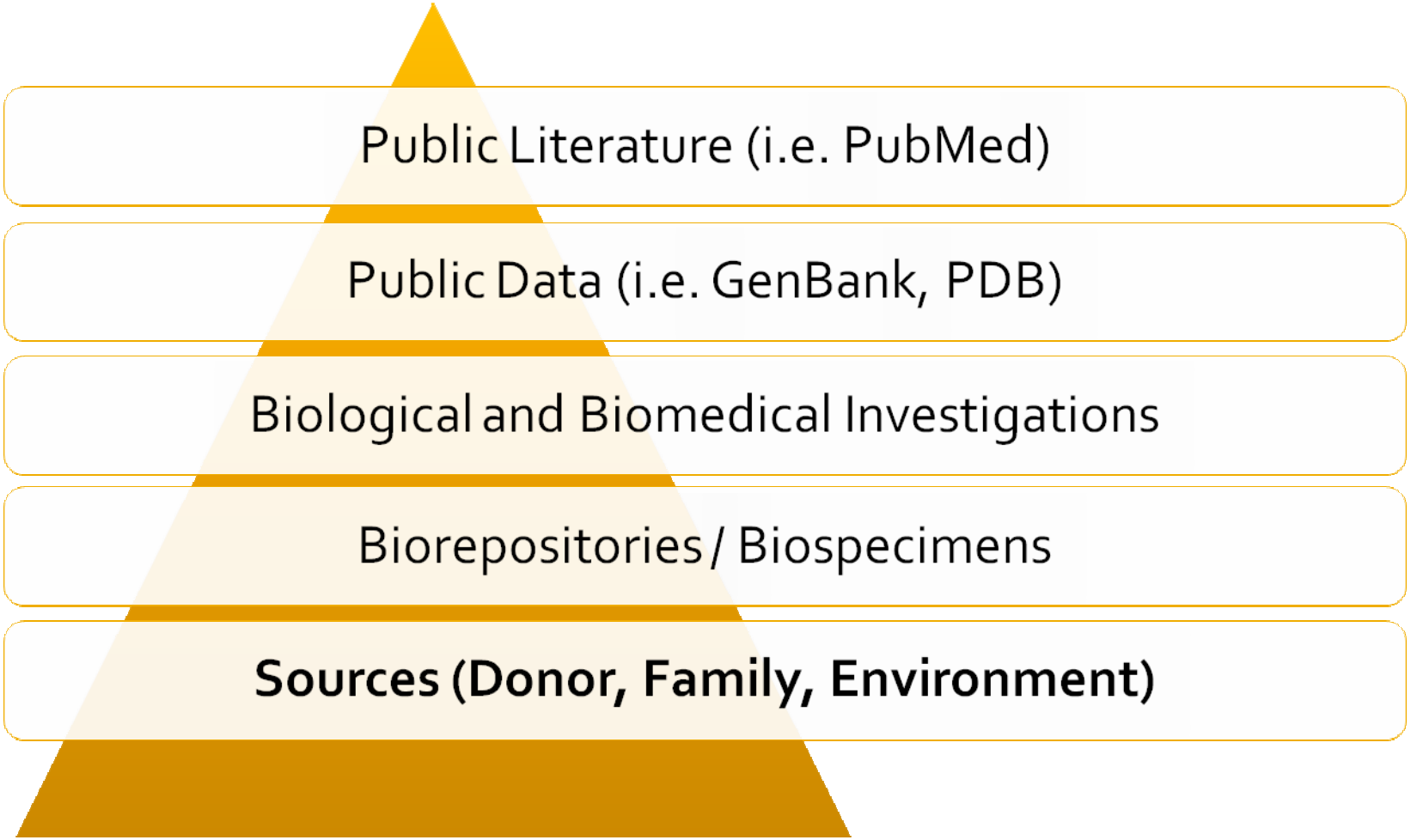
- From separation from source to analyte
- Containers, exposures, and "treatments"
- Unique identifiers / barcodes
- System users, locations and timestamps
- Security and audit trail
- Clinical pathology / laboratory medicine findings
  - Clinical tests and results & images (TRIs)
  - Quality assurance / quality control TRIs
  - Research TRIs

- **Retrospective** information (queries and reports)
- **Prospective** information (SOPs and workflow support)

Advanced Biorepository Systems

- Precise sample provenance tracking
  - Links back to source / donor
  - Links forward to investigations, data and publications
- Detailed collection, handling, processing data
  - From giant binders of SOPs to FuGE protocols
  - Supporting documents (i.e. consents, protocols)
- Other links, integration and interoperability
  - Uses publically documented interfaces (APIs)
  - Uses standard vocabularies / terminologies
- Stewardship information
  - Donor preferences, consents, authorizations
  - Organizational or regulatory constraints
- Tracking of variable costs

Informatics Framework Components



Information About the Source

- Human source / donor / patient
  - Individual identifiers, history, environment, family
  - Longitudinal follow-up after specimen collection
  - CDISC study data standards
  - Biomedical / health data standards (i.e. FMA, UMLS)
  - How many CDEs describe a person? 50? 1500?
- Non human source
  - Individual: identifiers, history, environment, family
  - Collection / processing location
  - Other relevant standards (i.e. Taxonomy, CARO)

Biorepository Data / Object Models

- **caTissue** UML model
- Common Biospecimen Model (**CBM**)
  - Specimen Locators
    - http://biospecimens.cancer.gov/locator (NCI)
    - http://biospecimens.ordr.info.nih.gov/ (NIH)
    - Harvard Catalyst
    - CHTN http://www.chtn.nci.nih.gov/biospecimens.html
    - Arizona Specimen Locator
- caLIMS2 (to connect caTissue with caArray, etc)
- **FuGE** UML model for investigations
  - Model for standard operating procedures (SOPs)
    - Processing steps
    - Operators
    - Timestamps
  - Extensible (e.g. FuGEFlow)
  - Could support biospecimen science

Data Exchange Formats

- XML
  - MAGE-ML
  - GelML
  - RDML (RT-PCR)
  - mzML (mass spectrometry)
- TAB / spreadsheet
  - MAGE-TAB
  - **ISA-TAB**

Controlled Vocabulary / Ontology

- UMLS Metathesaurus
  - NCI Thesaurus (**NCIt**)
    - caTissue
    - Common Biospecimen Model (CBM)
  - LOINC
  - SNOMED-CT
  - CPT, ICD, etc
- Ontology for Biomedical Investigations (**OBI**)

Biorepository System Options

5AM Solutions – Specimen Locator: http://www.5amsolutions.com/
Artificial Intelligence in Medicine - **TissueMetrix**: http://www.aim.on.ca/
Aptia – Visual Specimen Manager: http://www.aptia.com
BioFortis – **LabMatrix**: http://www.biofortis.com
Caisis: http://www.caisis.org
caTissue: http://catissuecore.wustl.edu/
Daedalus Software – BTM: http://www.daedalussoftware.com
Freezerworks: http://www.freezerworks.com/
GenoLogics – BioVault: http://www.genologics.com/
GenVault: http://www.genvault.com/
Healthcare IT – **BIGR** (formerly Ardais): http://www.healthcit.com
IMS **- BSI-II**: http://www.bsi-ii.com/
LabAnswer: http://www.labanswer.com
LabVantage - Sapphire: http://www.labvantage.com/
LabWare LIMS: http://www.labware.com
Ocimum Biosolutions - **Biotracker**: http://www3.ocimumbio.com/
PercipEnz – **OnCore**: http://www.percipenz.com/
PhaseForward – **Waban**: http://www.phaseforward.com/
Thermo Scientific – **Nautilus** LIMS: http://www.thermo.com

Biorepository Informatics Reality

- Many "workhorse" Excel and Access systems
- caTissue emerging as standard in cancer
- CTSAs/i2b2 developing repository informatics

- Need for practical informatics ecosystem
  - Moving towards the standards
  - Migration to enterprise solutions
  - Integration with data repositories, AP-LIS and CP-LIS
  - Public interfaces to CBM, caTissue, i2b2, etc
  - Data capture with simple workflow-friendly systems (i.e. AP-LIS, CP-LIS, REDCap)
  - Integrate biospecimen research with repository operations

Take Aways

- Avoid creating new biospecimen terminologies, formats, object models if an existing one can be adopted or adapted
- Align, harmonize and link data and systems across five components of framework
- Keep the costs and barriers to use low
  - Reality of long tail of investigators / labs
  - Reality of time and resource pressure/constraints
- Facilitate biospecimen science by integrating research with repository operations

Acknowledgements

- University of Washington
  - Nicholas Anderson
  - Malia Fullerton
  - Kelly Fryer-Edwards
  - James Brinkley
  - Peter Tarczy-Hornoch
  - Lawrence True
  - Rodney Schmidt
  - David Chou
  - Marc Provence
- UCLA
  - Robert Dennis
  - Andrew Helsley
- Daedalus
  - Azita Sharif
  - Stefano Santoro
- ISB
  - Eric Deutsch
- ISBER Informatics WG
  - Cheryl Michaels, Freezerworks
- Prostate SPORE Informatics Group
  - UW, UCLA, UCSF
  - Baylor, MDACC
  - Northwestern, Mayo Clinic, U of Michigan
  - Harvard/DFCI, Johns Hopkins, MSKCC
- NCI OBBR

Informatics Framework Components