Integrity ★ Service ★ Excellence

Building an Operational Data Hub with HyperThought™ and MarkLogic®

4/20/2018

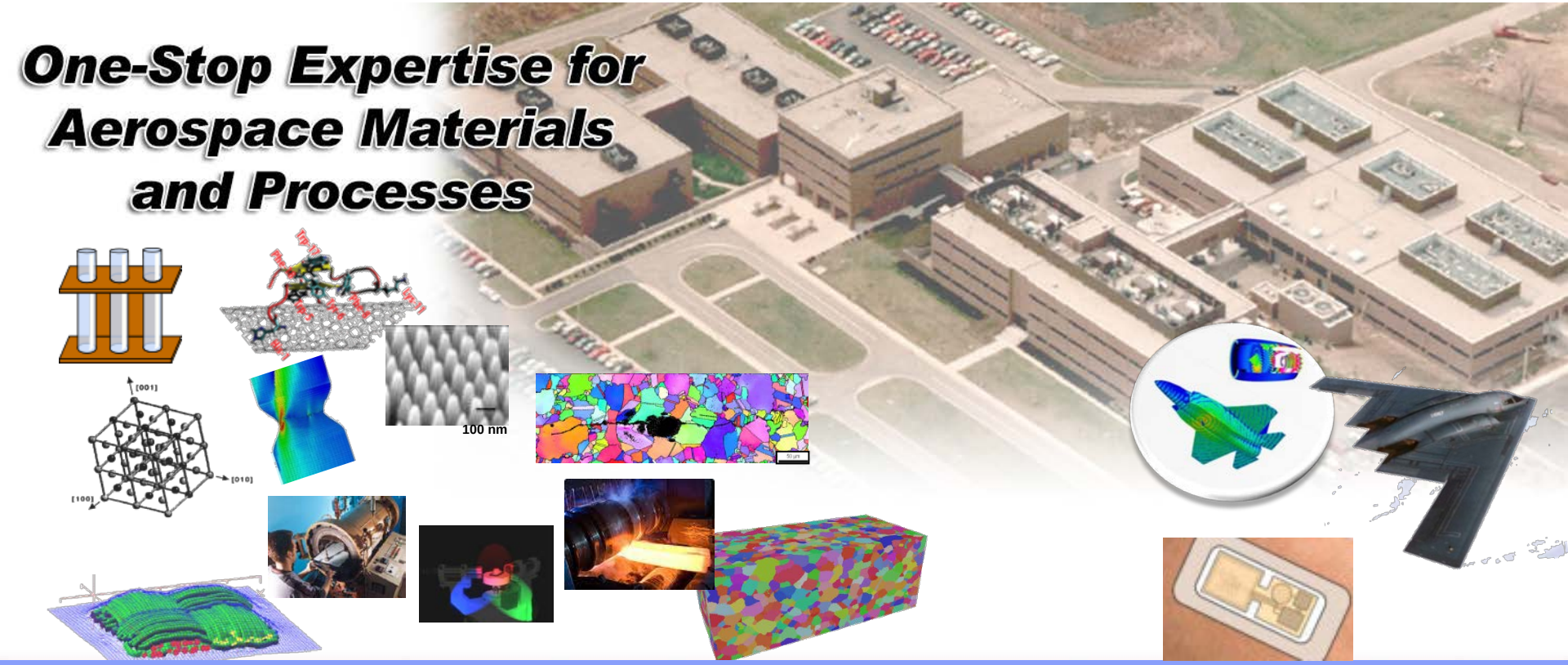
Matthew Jacobsen

Air Force Research Laboratory

Brent Perry

MarkLogic

Air Force Research Laboratory
Materials and Manufacturing Directorate



A full spectrum materials & manufacturing organization:

Metals / Ceramics / Composites / NDE / Semiconductors / Polymers / Photonic Materials / Biomaterials

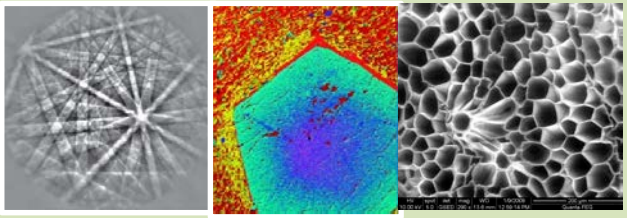
Structural / Propulsion / Weapons / Sensors / Survivability Applications

Discover... Design... Manufacture... Transition... Support

Materials and Manufacturing
Research Infrastructure

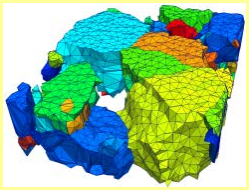
- 700+ scientists and engineers
- 108,000 sq ft lab space, 200 lab modules
- 750+ computers associated with research equipment
- 1000+ computers on desks: 2 separate networks
- 80+ scientific and engineering software packages
- Local computational clusters & remote HPC

Materials Characterization Facility



DSRC Lightning

High Energy Diffraction Microscopy



And no supporting collaborative research environment

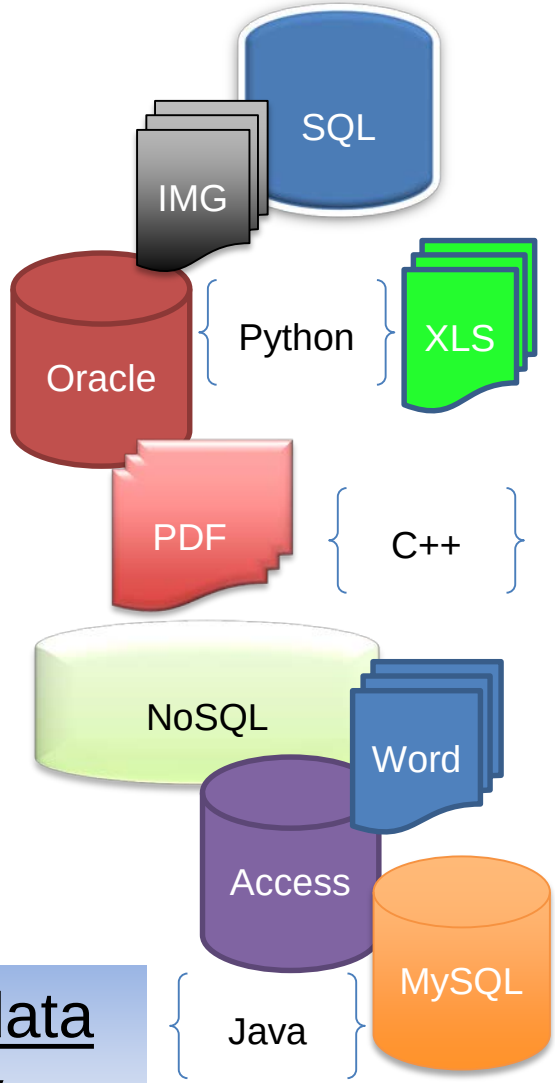
Problem Space

Unmanageable Data:

- Volume – exabyte quantity
- Velocity – millisecond demand
- Variety – multi-model definition
- Veracity – full pedigree

Numerous Failure Modes:

- Data are poorly described and held
- Software is inflexible and inaccessible
- Policy is outdated – no governance



Model-driven research limited by disparate data sources, little coordination, no reproducibility

MarkLogic/AFRL Partnership

- AFRL was the first US Research Lab partner
  - Visionary team, but organization resources for research more than IT
  - Atypical bottom-up approach to data fusion and ontology
  - Significant data challenges from multitude of parallel fascinating simulations, materials testing, and data modeling activities
  - Tremendous opportunity to boost efficiencies through data management
- Ingredients for Success
  - Great informal working relationship
  - Face-time at the whiteboard
  - Guiding ideas from concept to execution

Solution in Concept

Deploy a data management system for the research community that implements an Operational Data Hub

- Must be iteratively developed according to agile principles
- Must be highly flexible
- Must be scalable and deployable
- Must allow build vs. buy decisions for components
- Must allow for hybrid storage/connection model
- Must support model-based domain definitions

Enter MarkLogic

Problem: Early efforts in NoSQL (Mongo) held promise, but required major effort to tether together and exploit. Needed to add search, security, semantics, and performance at scale.

Solution: An Operational Data Hub provides high speed access to both centralized and co-located data.

- Multi-model context to cut through complex data diversity
- MarkLogic provides complete picture of data
- Data-linking supported via the semantic index
- Metric: 100x increase in access performance over legacy relational approach

MarkLogic Value Proposition

Feature Support

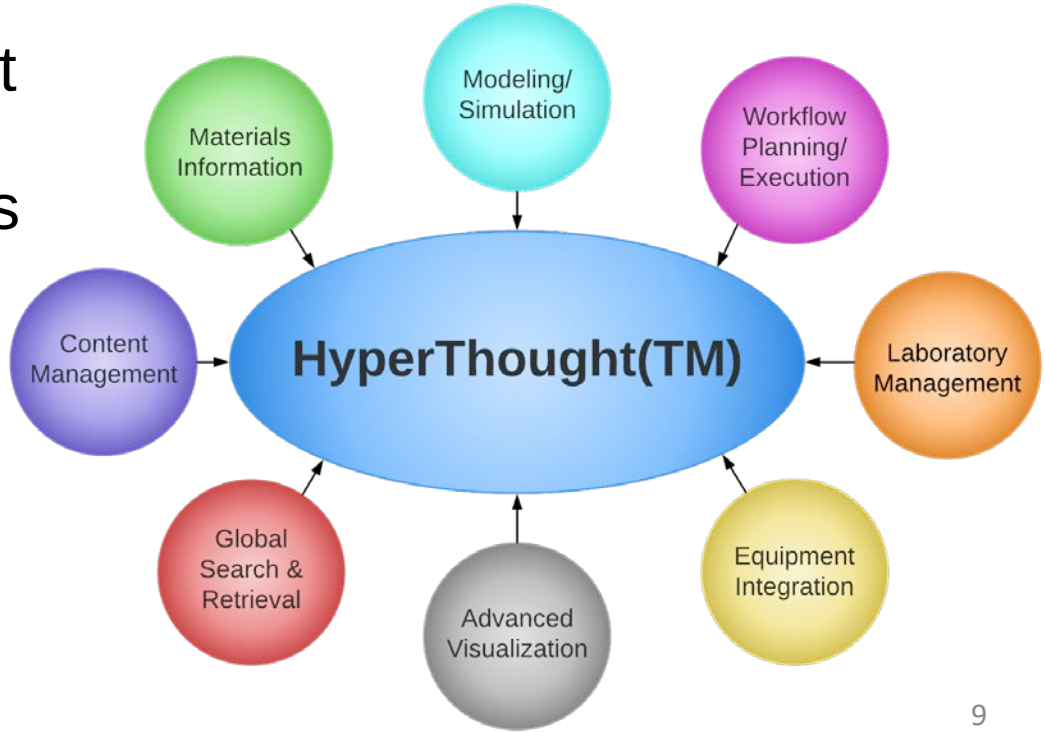
- Search and Discovery within high volume, velocity and variety data sets
- Semantic capabilities are required to unpack disparate data (repos are distributed over dozens of silos)
- Schema-agnostic NoSQL is also required to accommodate rapidly-changing multi-model entities

What about Open-Source Software?

- Maintaining open source solutions would require more than a half dozen hand-stitched components
- Tens of thousands of hours of development saved

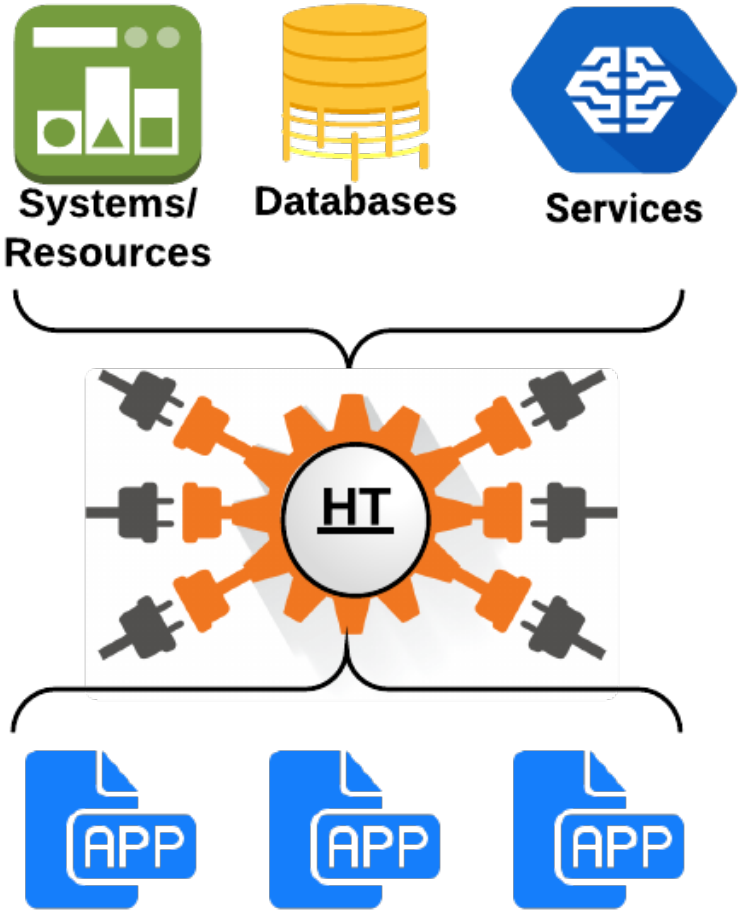
HyperThought™

- Combines best-of-breed commercial and open-source tools
- Avoids a single development model (monolithic, FOSS, etc.)
- Paves the way for a true multi-model representation
- Success requires a joint effort between software and materials engineers to deliver game-changing functionality



Two Pillars

- Integration layer – interface for connecting and orchestrating systems, along with core collections of micro-services
- Tool suite – user experience including
  - content management
  - digital workspaces
  - equipment integration
  - workflow management
  - visualization tool
  - search engine



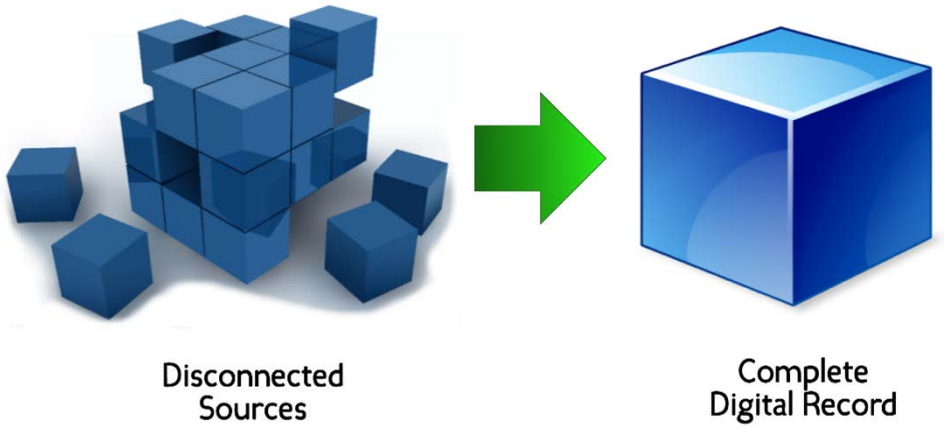
HyperThought™

Integration of Capabilities and Data

- Complete experimental and computational material pedigree
- Robust digital workspaces for collaboration
- Integration with existing systems and equipment
- Data management for files, metadata, and datasets

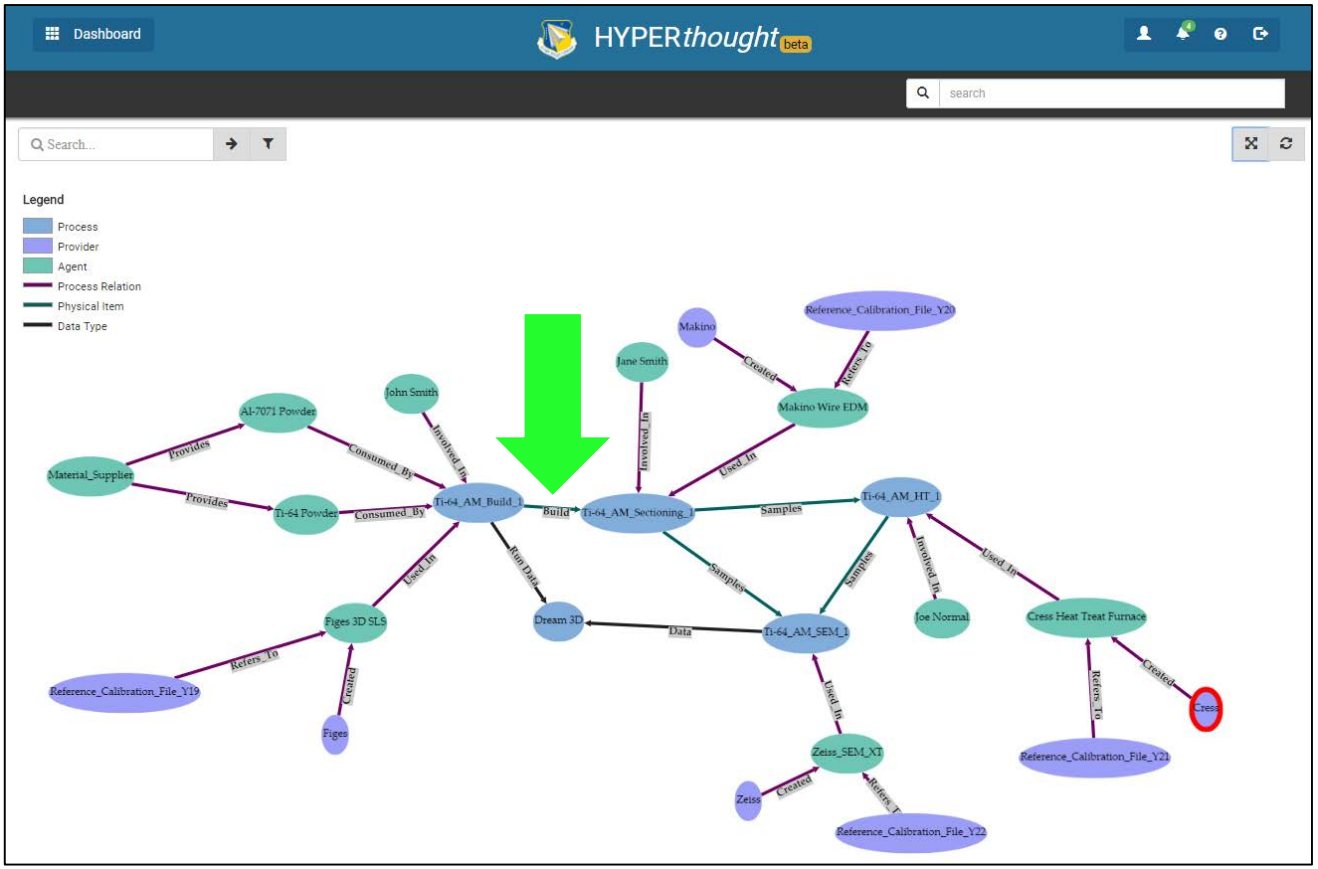
Cross-Domain Applications

- Technology-agnostic integration layer
- Modular and scalable architecture
- Inter-organizational connectivity
- User-tailored to any domain – ICMSE, Digital Thread, maintenance, etc.



Data Discovery via Pedigree

"Display all entities that have acted on specimen X"



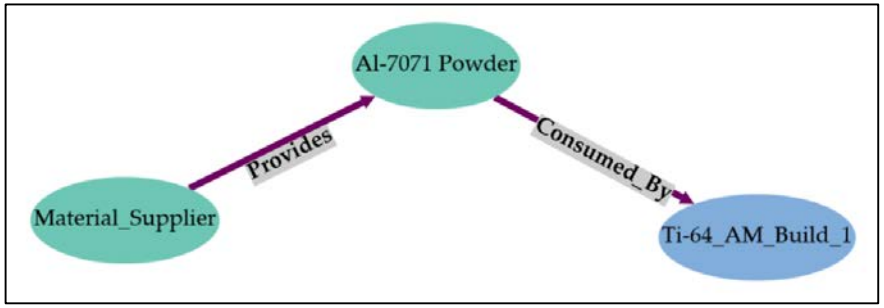
Data Discovery via Attribute Search

"Show all samples with 5 mm thickness and 30-degree print angle"

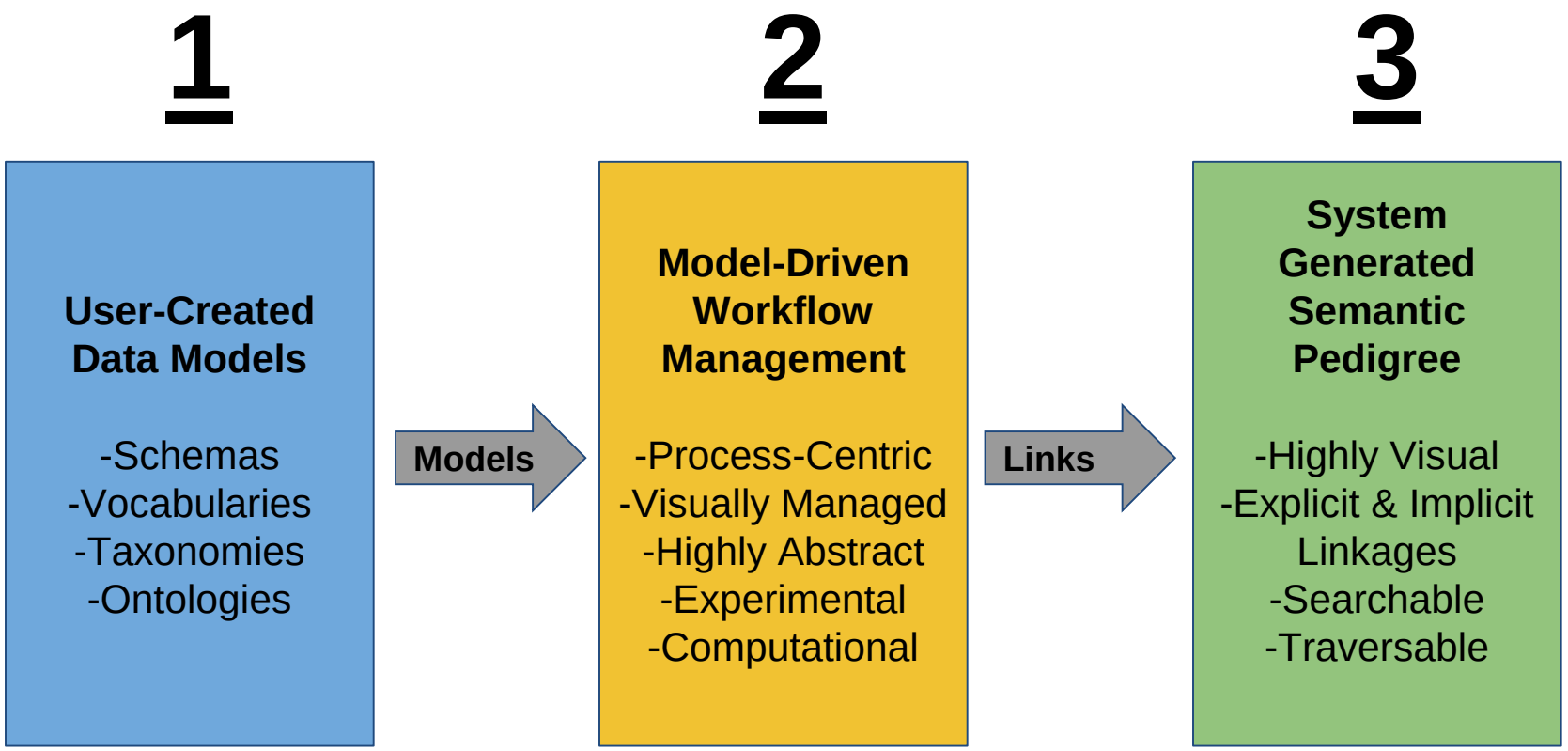
Sample with 5 mm thickness

Search

Object	Object Type	Date	Creator	Source
3D Printing	Project	2/8/2017, 12:00:00 AM	Lance Wilhelm	ICE
MarkLogic Prep	Project	11/8/2016, 12:00:00 AM	Matthew Jacobsen	ICE
+ MakertoMfgAnalysis.pptx	File	7/19/2017, 5:26:12 PM	Emily Fehrman-Cory	ICE
Prussa 3D Printer	Project	8/28/2017, 12:00:00 AM	Robert Lee	ICE
M-Hub-spool holdahMK3.stl	File	8/28/2017, 8:23:05 PM	Chad Meyer	ICE
mp-804	MaterialsProject	Invalid Date	Unknown or undefined	Materials Project
mp-830	MaterialsProject	Invalid Date	Unknown or undefined	Materials Project
mp-1007824	MaterialsProject	Invalid Date	Unknown or undefined	Materials Project
mp-2853	MaterialsProject	Invalid Date	Unknown or undefined	Materials Project

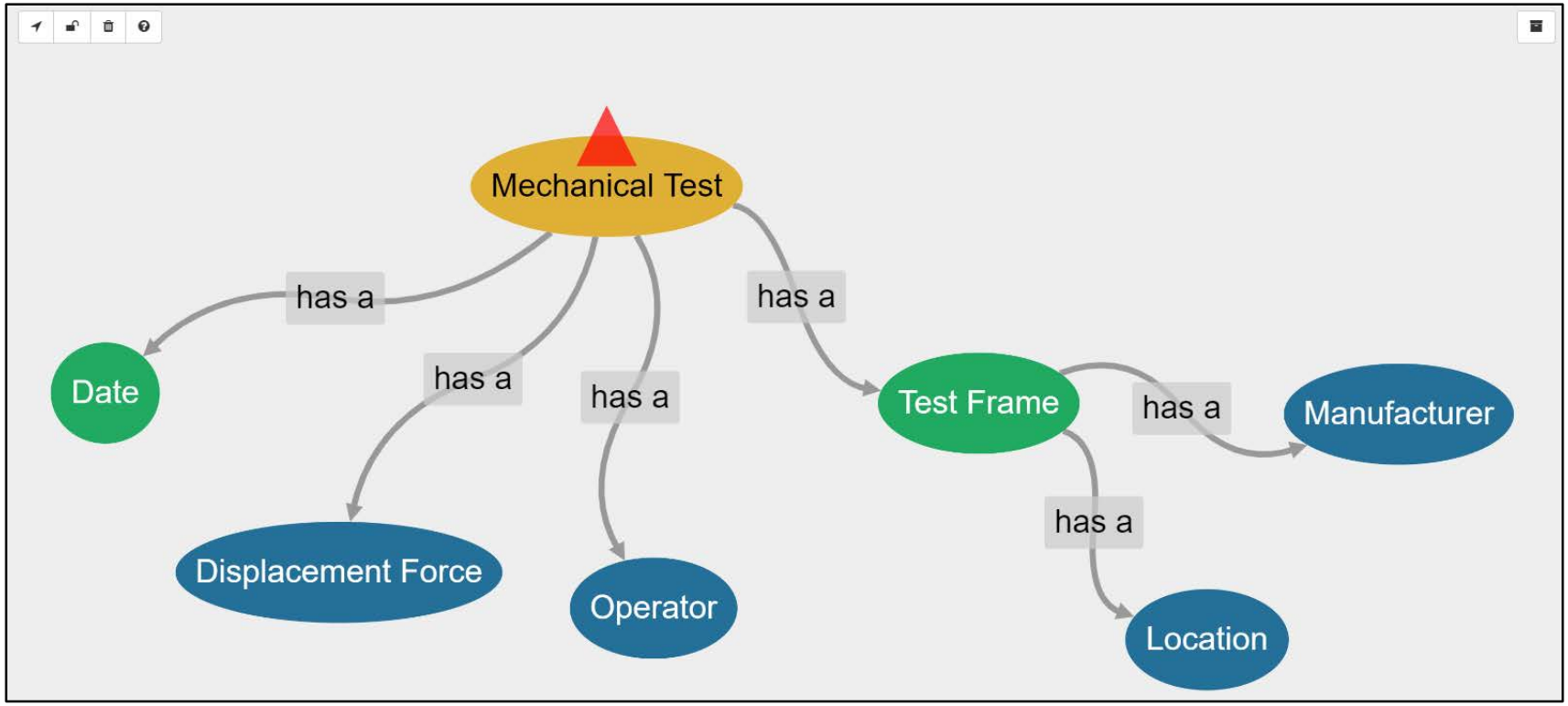


The Road to Discovery



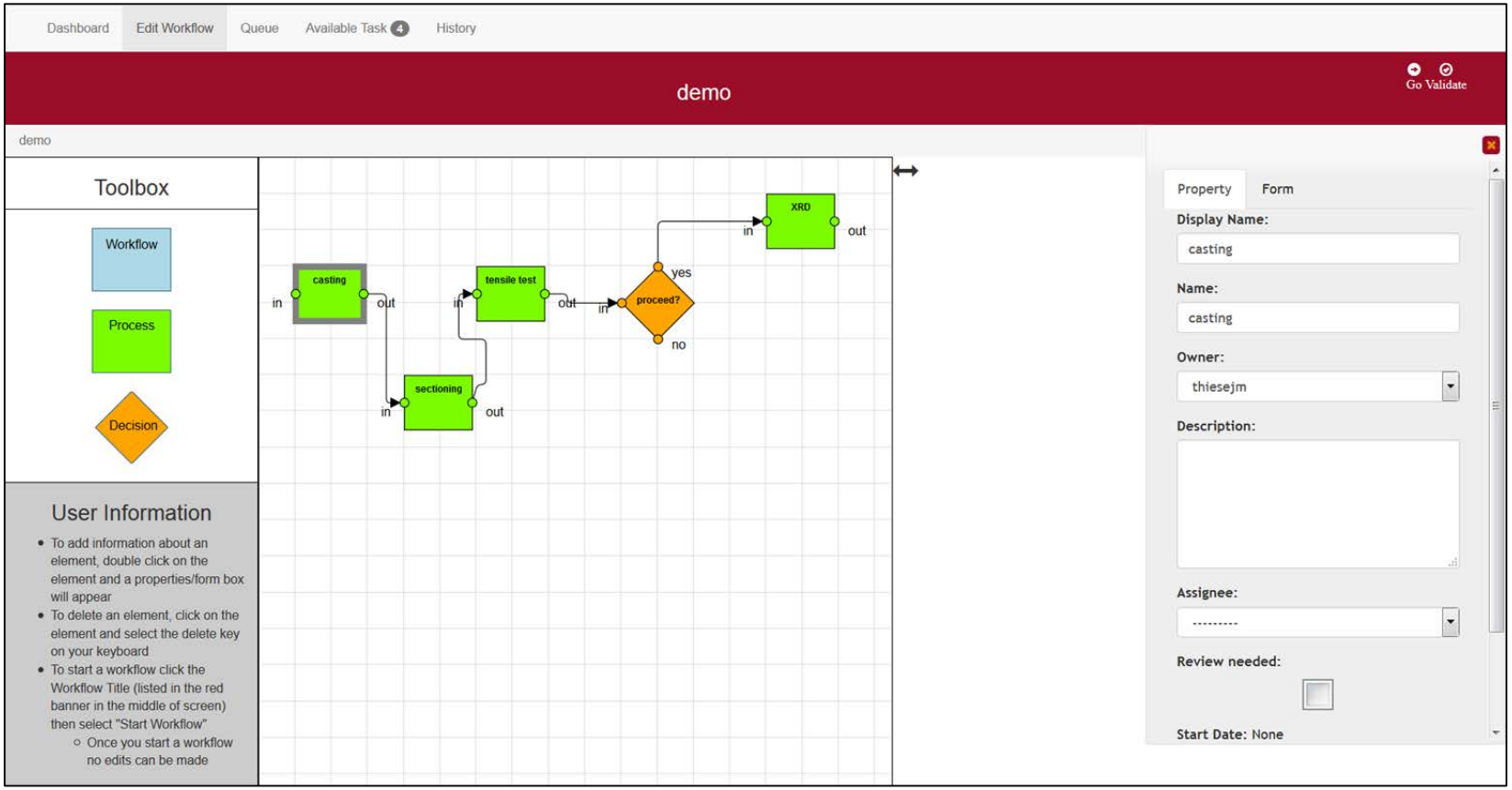
1. User-Created Data Models

- Users (not developers) define all aspects of their domain, to any level of desired granularity, as “Data Models”
- “Data Models” populate reusable templates, vocabularies, and full ontologies
- MarkLogic enables multi-model representation across enterprise components



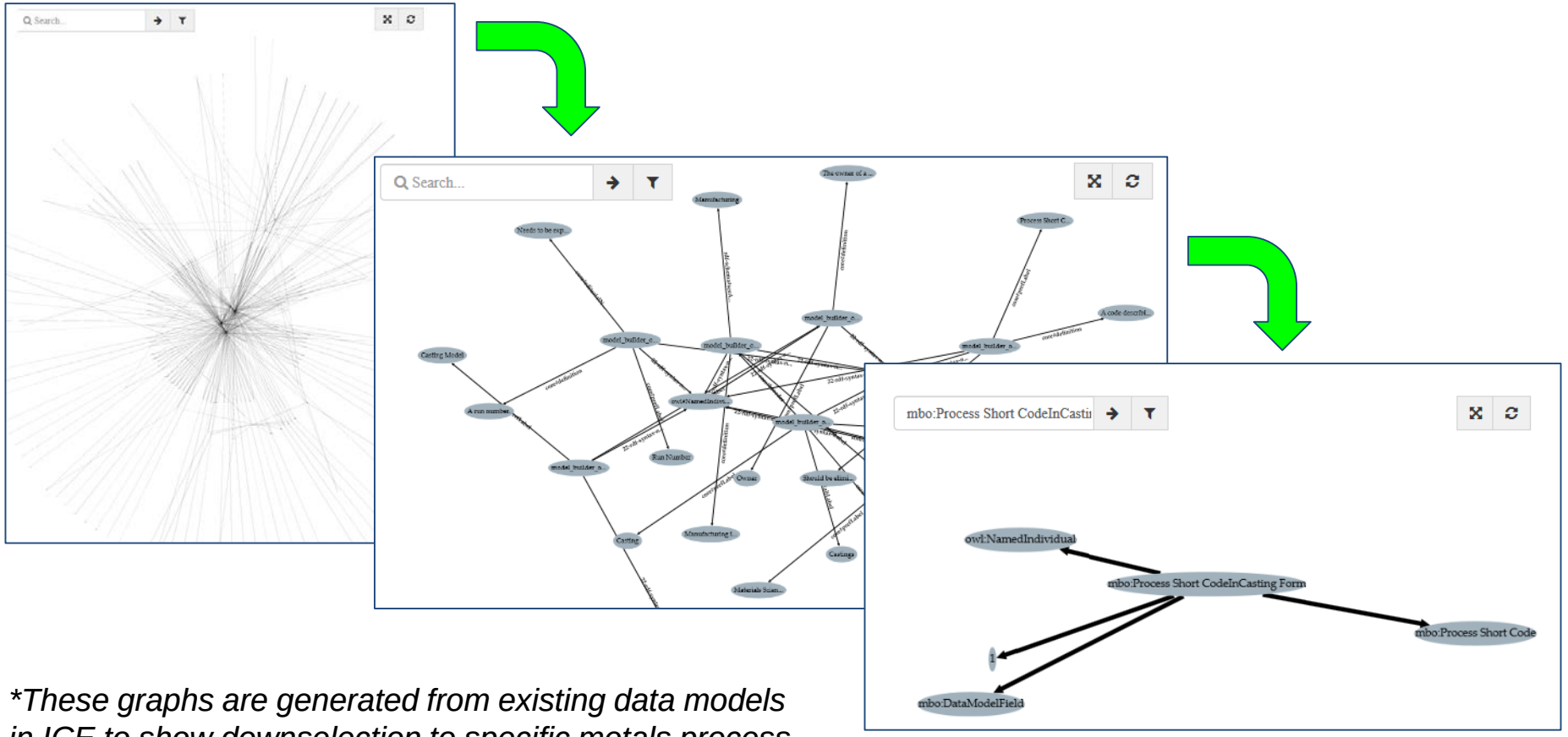
2. Model-Driven Workflow Management

- User-defined Data Models are linked together to create complex workflows
- Data forms, equipment data streams, and computational jobs can all be linked to capture entire research efforts (again requiring NoSQL flexibility)



3. System Generated Semantic Pedigree

- Workflow activities, data models, users, equipment, and other resources linked together to show highly scalable visual pedigree
- User-driven search and traversal of graph structures



**These graphs are generated from existing data models in ICE to show downselection to specific metals process short-codes*

HyperThought with MarkLogic Value Proposition

Maximize the return on investment in research by:

- Preventing redundancy
- Optimizing research resources
- Leveraging collaborative efforts

Key Principle – FAIR:

- Findable
- Accessible
- Interoperable
- Reusable

Next Steps - Additive

- Nation-wide challenge in AM exploration for metals
- Variety - every process flow, database, and object repo that is submitted is unique
- Volume – hundreds of terabytes for 12 month challenge set
- Veracity – micro-structural fidelity for data searching

PM: “We couldn’t do this without HyperThought/MarkLogic”

Next Steps – Rotor Data

- 40 years of weapons system test data in paper files
- Active aircraft rely on millions of documents in varying formats
- Scan -> OCR -> index/facet -> search
- Weeks of effort down to seconds
- Data reflect 2.6B in operational materials

PM: “Any effort of this scale is a game changer in how we do life-ing/predictive modeling”

Summary

- HyperThought has accelerated with MarkLogic:
  - Enterprise features reducing custom development needs
  - Raw performance and scaling
- AFRL-RX success with HyperThought has spurred interest across Directorates and Labs
- Upcoming Use Cases
  - Rotor data, analytics with Ayasdi, SCRAMJET particle simulations