CGIAR Consortium

CGIAR is a global research partnership for a food secure future

CGIAR Open Access and Open Data – Overview

Medha Devare

(m.devare@cgiar.org)

Arid Cereals, Minneapolis

November 14, 2015

CGIAR R4D Centers

Why OA/OD??

CGIAR's decades-rich data trove – availability for value addition via re-use, meta-analyses, decision support…??

- Enhance cross-regional, cross-disciplinary learning
  - Multiple CG centers, partners, countries, hubs
  - Varied data streams (breeding, agronomy…)
- Facilitate internal/external monitoring and evaluation
- Cement institutional memory
- Increase efficiency, RoI: Donor, public, scientist…
- CGIAR outputs are public goods

IMPACT…not impact factor!

Recurrent OA/OD-related themes

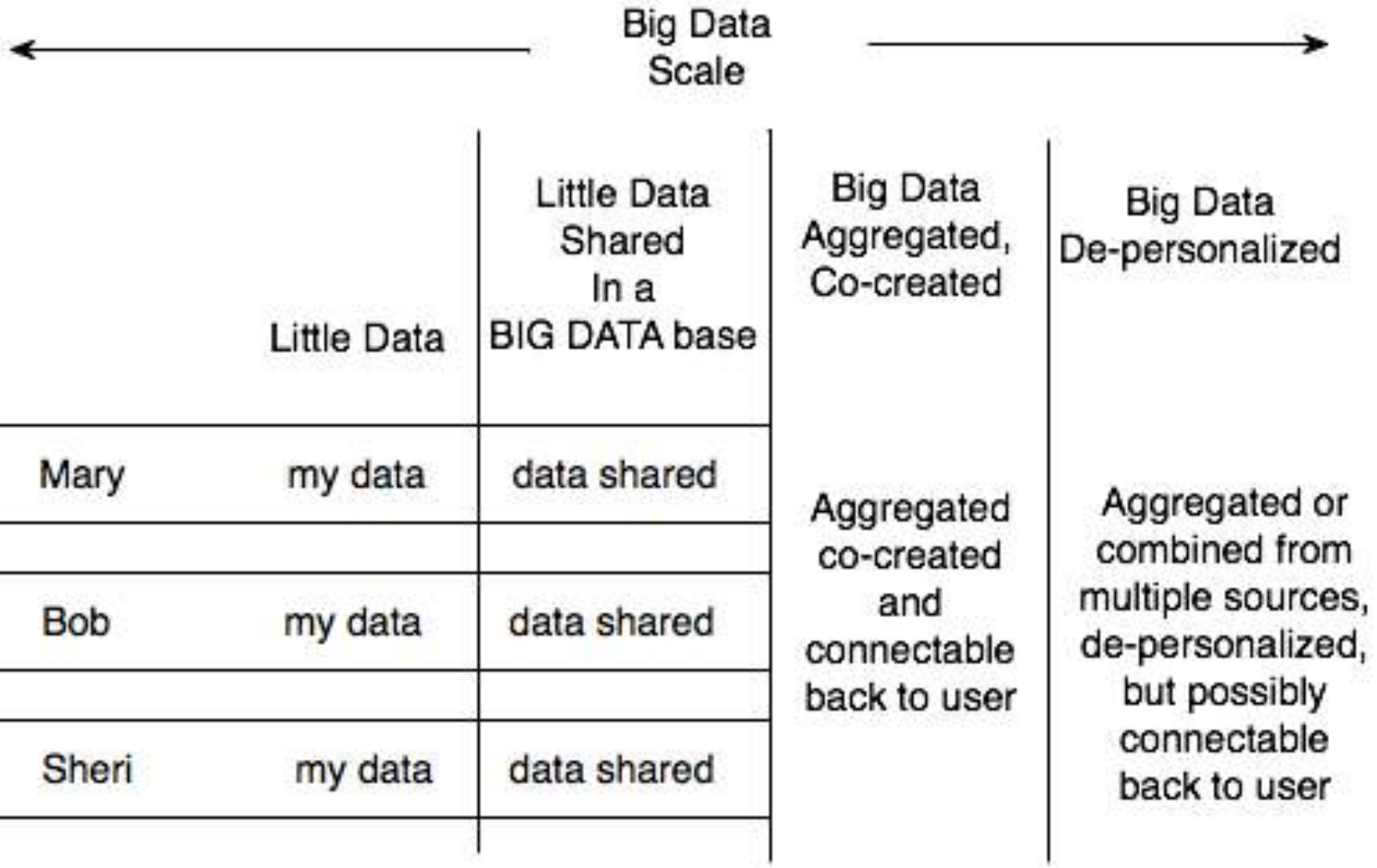
Need to validate with data…

Sustained info services…

Big data and ICTs as an opportunity…

Variety of data needed, in context, aggregated…

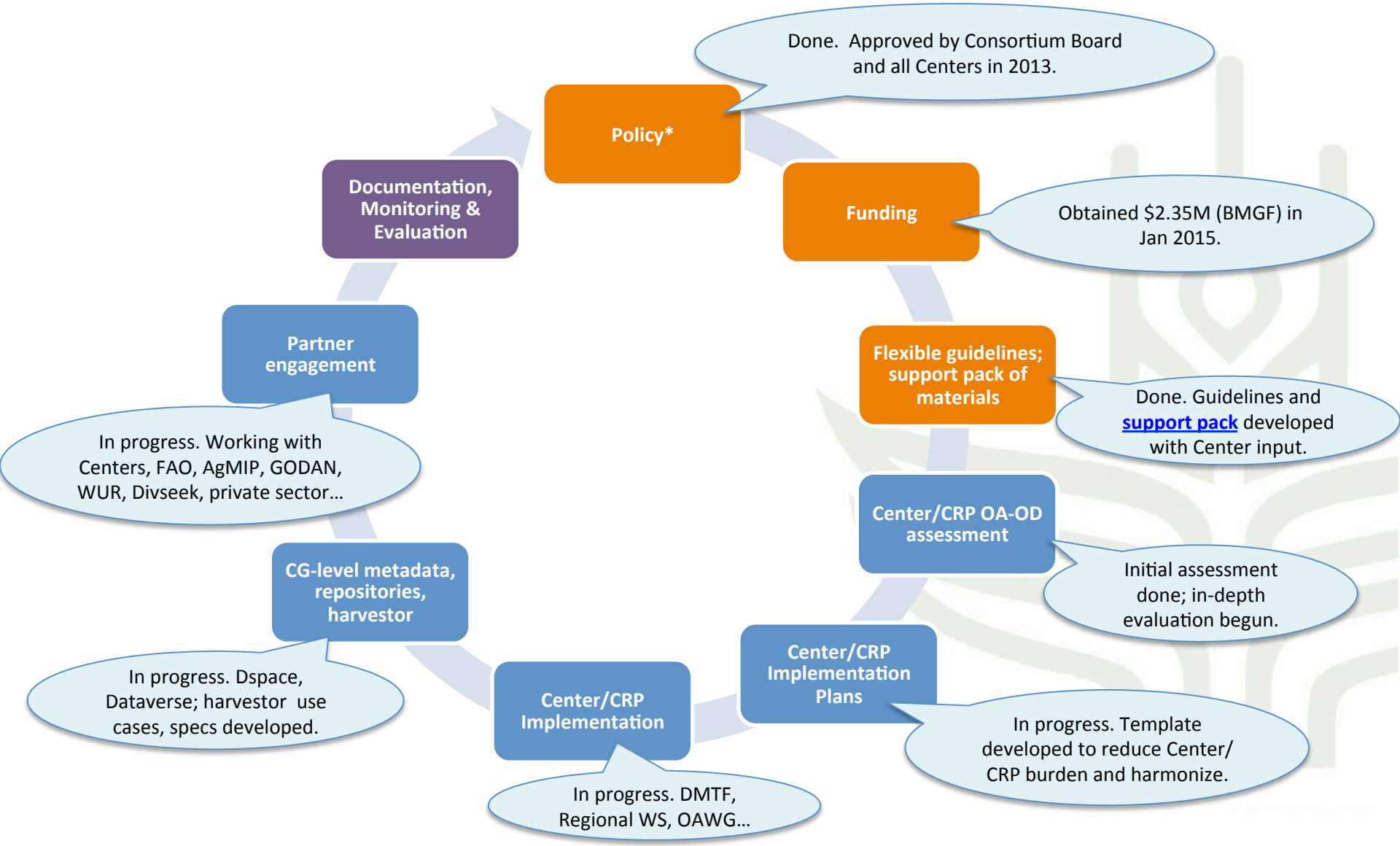
Access to high quality, plug ‘n play data…



http://napsterization.org/stories/archives/cat_personal_data.html

BIG DATA HIGHWAY?

CGIAR OA-OD: Overview



*Final policy available at www.cgiar.org/open

What do we need to accomplish?

Get from this…

http://en.wikipedia.org/wiki/File:Ralls_Texas_Grain_Silos_2010.jpg

To this:

Which requires access to data sets and harmonization on…

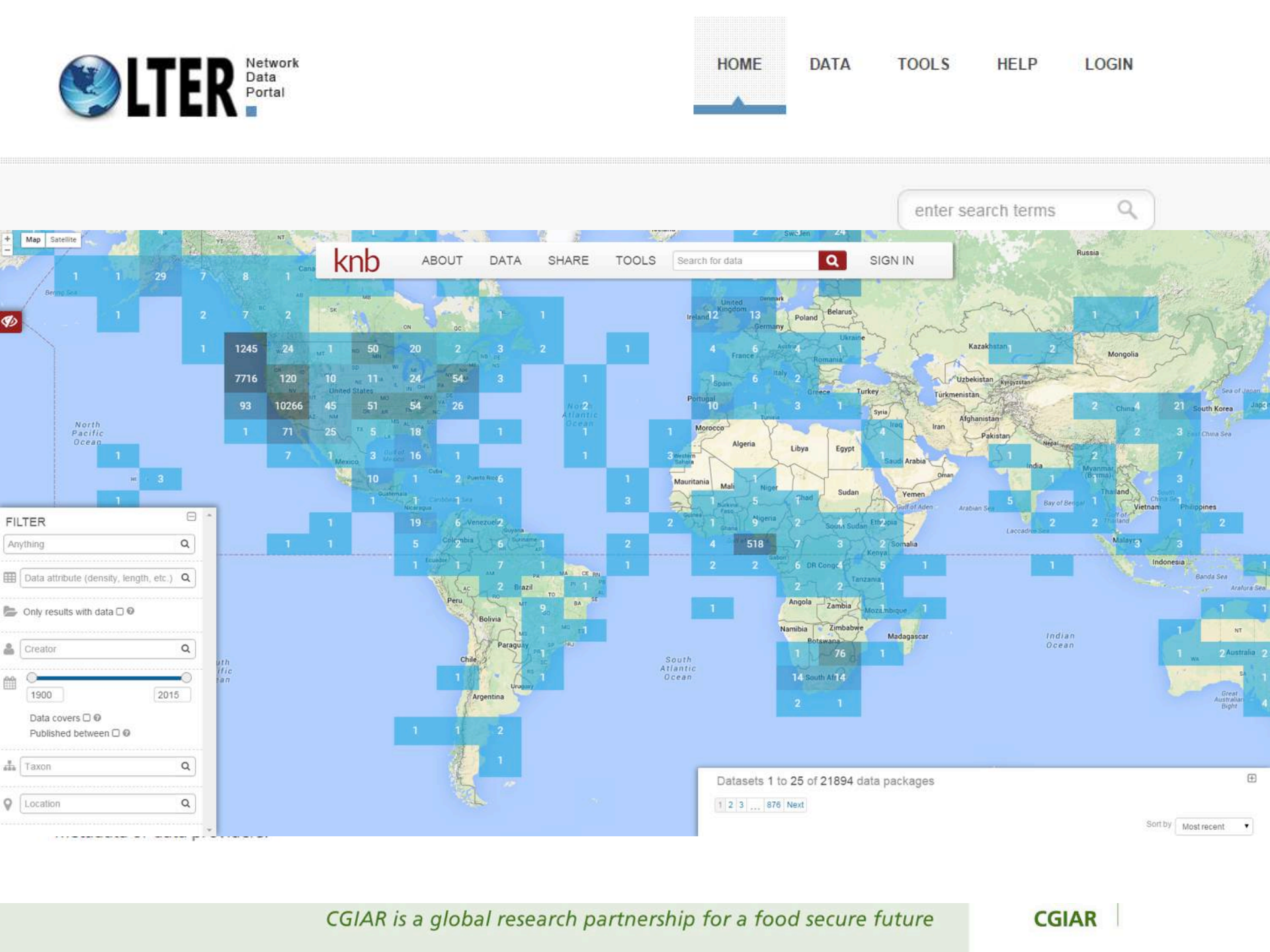
standards

interoperability

metadata

incentives/culture

tools/platforms


LTER Network Data Portal
HOME
DATA
TOOLS
HELP
LOGIN
enter search terms
knb
ABOUT
DATA
SHARE
TOOLS
Search for data
SIGN IN
FILTER
Anything
Data attribute (density, length, etc.)
Only results with data
Creator
1900
2015
Data covers
Published between
Taxon
Location
Datasets 1 to 25 of 21894 data packages
1 2 3 ... 876 Next
Sort by
Most recent
CGIAR is a global research partnership for a food secure future
CGIAR

The bigger picture…

Display Settings: GenBank

Send:

Zea mays cultivar B73 chromosome 4

GenBank: CM000780.3

FASTA Graphics

Go to:

```
LOCUS       CM000780           242029974 bp    DNA     linear   CON 24-OCT-2013
DEFINITION  Zea mays cultivar B73 chromosome 4.
ACCESSION   CM000780 JH967973 JH967981 JH967982
VERSION     CM000780.3  GI:552562410
DBLINK      BioProject: PRJNA10769
KEYWORDS    .
SOURCE      Zea mays
  ORGANISM  Zea mays
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; PACMAD
            clade; Panicoideae; Andropogoneae; Zea.
REFERENCE   1  (bases 1 to 242029974)
  AUTHORS   Schnable,P.S., Ware,D., Fulton,R.S., Stein,J.C., Wei,F.,
            Pasternak,S., Liang,C., Zhang,J., Fulton,L., Graves,T.A., Minx,P.,
            Reily,A.D., Courtney,L., Kruchowski,S.S., Tomlinson,C., Strong,C.,
            Delehaunty,K., Fronick,C., Courtney,B., Rock,S.M., Belter,E.,
            Du,F., Kim,K., Abbott,R.M., Cotton,M., Levy,A., Marchetto,P.,
            Ochoa,K., Jackson,S.M., Gillam,B., Chen,W., Yan,L.,
            Higginbotham,J., Cardenas,M., Waligorski,J., Applebaum,E.,
            Phelps,L., Falcone,J., Kanchi,K., Thane,T., Scimone,A., Thane,N.,
            Henke,J., Wang,T., Ruppert,J., Shah,N., Rotter,K., Hodges,J.,
            Ingenthron,E., Cordes,M., Kohlberg,S., Sgro,J., Delgado,B.,
            Mead,K., Chinwalla,A., Leonard,S., Crouse,K., Collura,K.,
            Kudrna,D., Currie,J., He,R., Angelova,A., Rajasekar,S., Mueller,T.,
            Lomeli,R., Scara,G., Ko,A., Delaney,K., Wissotski,M., Lopez,G.,
            Campos,D., Braidotti,M., Ashley,E., Golser,W., Kim,H., Lee,S.,
            Lin,J., Dujmic,Z., Kim,W., Talag,J., Zuccolo,A., Fan,C.,
            Sebastian,A., Kramer,M., Spiegel,L., Nascimento,L., Zutavern,T.,
            Miller,B., Ambroise,C., Muller,S., Spooner,W., Narechania,A.,
            Ren,L., Wei,S., Kumari,S., Faga,B., Levy,M.J., McMahan,L., Van
            Buren,P., Vaughn,M.W., Ying,K., Yeh,C.T., Emrich,S.J., Jia,Y.,
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

LinkOut to external resources

Order P(M1) cDNA clones [OriGene]

Order protein 2C (NTPase) cDNA clones [OriGene]

Related information

Assembly

BioProject

Component Of

Components (Core)

Gene

GeneView in dbSNP

Genome

Identical RefSeq

Protein

PubMed

PubMed (Weighted)

Taxonomy

The bigger picture…

Resources How To

Search databases

Help

drought tolerance maize Search

Results found in 17 databases for "drought tolerance maize"

Literature

Books	28	books and reports
Policy briefs	0	policy briefs
Training, extension	0	books, journals, and more
AgPub	170	abstracts and citations
AgPub Central	4,580	full-text journal articles

Geographies

Asia	30	resources for sub-regions and countries in Asia
Africa	82	resources for sub-regions and countries in Africa
Middle East	11	resources for countries in the Middle East
Europe	13	resources for sub-regions and countries in Europe
N. America	17	resources for sub-regions and countries in N. America
S. America	25	resources for sub-regions and countries in S. America

Toolkit

Technology tracker	0	technology adoption tracking
Agri-semantics	5	ontologies, vocabularies
AMKN	0	climate change adaptation and mitigation knowledge network
Activity mapper	2	project and activity mapping
Methods	42	methodologies

Subjects

Agroforestry	3	resources related to agroforestry
Agronomy	22	resources related to agronomy
Aquaculture/fisheries	0	resources related to aquaculture/fisheries
Climate change	155	resources related to climate change
GIS/remote sensing	33	resources related to GIS/remote sensing
Genebank	53	genebank resources
Genetic/genomic	5	genetic/genomic resources
Hydrology/water mgmt.	0	resources related to hydrology/water management
Livestock/animal breeding	120	resources related to livestock/animal breeding
Natural resource mgmt.	60	resources related to natural resource management
Plant breeding	134	resources related to plant breeding
Plant protection	13	resources related to plant protection
Socioeconomics/livelihoods	47	resources related to socioeconomics/livelihoods
Other	56	resources related to other subjects

Aligned CG-donor policies, guidelines

USAID, Gates, DFID, ACIAR…
Definitions (what is OA?), consequences (e.g. 80-20 funds)
Data mgmt. plans, implementation templates, indicators
publishing/licensing guidelines
Institutional or community interoperable repositories
OA-OD governance/org…

Aligned CGIAR plans, approaches

- Overall: implementation plans
- HR: evaluations, KPIs ↔ ODjAR
- HR: contractual language, induction process
- Legal: CC licensing
- Leadership: overall buy-in
- DM/KM: infrastr, support, guidance, messages
- ICT: infrastructure, support
- Partnerships: FAO, CABI, GODAN, AgMIP, Alterra-WUR, DivSeek, INRA, CIRAD…

OA-OD capacity, buy-in

CGIAR++: overall buy-in | Partners | Govts

begun, may be continued in phase II
phase II

Infrastructure

Analytics/tools

- Technology mapping - foresight
- Investment mapping - foresight
- M+E
- Research discovery
- Decision support

Interoperability

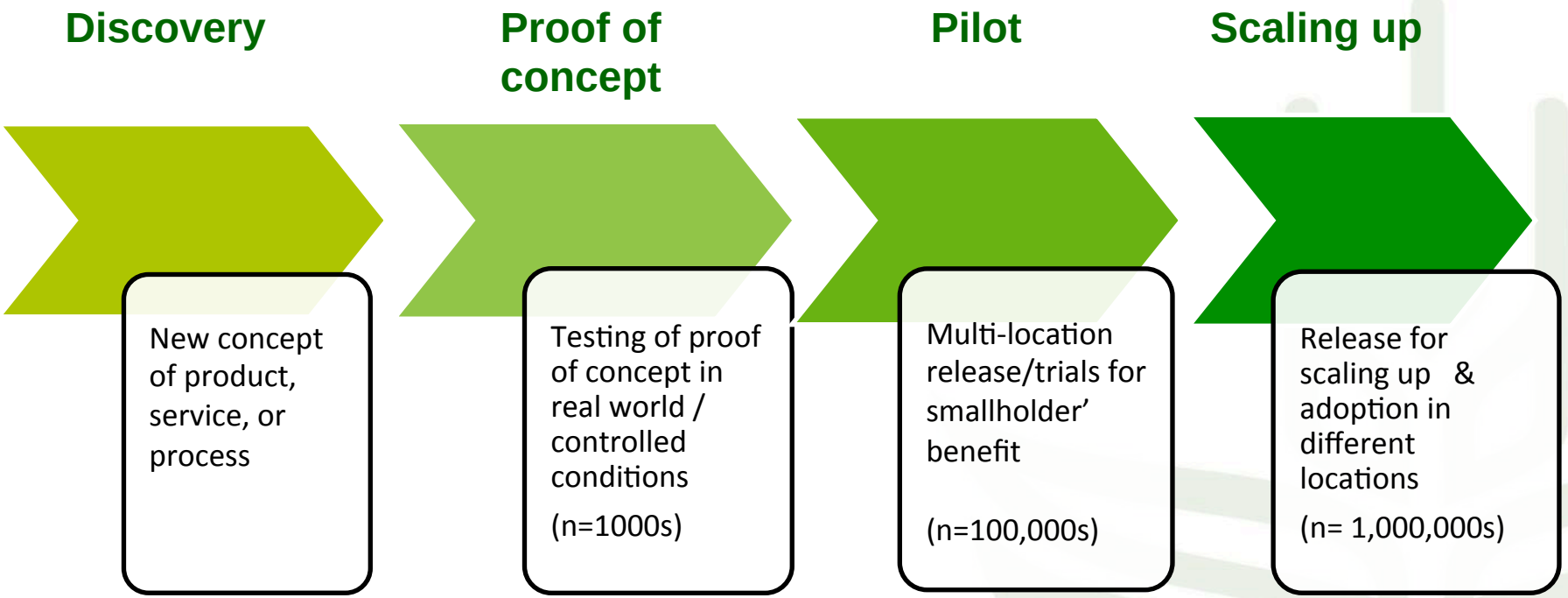
- Repositories
  standards, protocols
  CG core metadata
- Ontologies
  crop – agron – agritech – livest – value chain – agriVIVO
- Vocabularies
  GACS (AGROVOC-NAL-CAB)
- Geospatial
  Standards, metadata

Harmonized data sets

- BMS ↔ Agronomy field book
  ICASA – MD data dictionary
  ontology interlinkages
- Data mgmt. SOPs
  donor-aligned?
- Standard techs
  technology cat
- Geospatial
  consistency, quality

Thanks

General approach to R4D



Adapted from: P. Ellul

DATA??

CGIAR OA-OD project: Key objectives

1. Conduct a broad inventory and assessment of CGIAR capacity in OA-OD.

   Output 1.1. In-depth analysis of the publications landscape across CGIAR. → in progress

   Output 1.2. Needs analysis of CGIAR data management and data quality practices over the data life cycle. → in progress

   Output 1.3. Analysis of other research products across CGIAR. → in progress

   Output 1.4. Identification of gaps in CGIAR human infrastructure and enabling environment for OA-OD. → in progress

   Output 1.5. Proof-of-concept indexing tool/portal. → exploration

CGIAR OA-OD proposal: Key objectives

2. Develop a legacy data prioritization framework.

   Output 2.1. Data prioritization framework.

CGIAR OA-OD project: Key objectives

3. Provide coordinated support to Centers and CRPs in their efforts towards OA-OD, and leadership for external efforts.

Output 3.1. Support pack for Open Access and Open Data.

→ *Support pack v.1 in place*

Output 3.2. Consistent OA-OD implementation plans.

→ in progress

Output 3.3. Improved, interlinked initiatives, tools, and platforms.

→ Ongoing: Agronomy Ontology, AgTrials, IBP-AMS, AATP-VIP

Output 3.4. OA-OD collaboration and leadership beyond CGIAR.

→ AgMIP, agrisemantics harmonization (FAO, CABI, NAL+), WUR, INRA, CIRAD, GODAN…

CGIAR OA-OD proposal: Key objectives

4. Plan for impact assessment.

   Output 4.1. OA-OD impact assessment framework.

5. Plan for Phase 2: Implementation.

   Output 5.1. Phase II funding proposal.

3.4 Data storage and preservation

- Should raw data be preserved, or processed/normalized/ transformed data?
- Description of data to make it relevant in the future?
- Where will data be preserved? Is that location stable?
- Security mechanisms, backup procedures?
- Which file formats are best for long-term preservation?

Data storage and preservation – file formats

Type of data	Acceptable formats for sharing, reuse and preservation	Other acceptable formats for data preservation
Quantitative tabular data with extensive metadata **a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data**	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information some structured text or mark-up file containing metadata information, e.g. DDI XML file	proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta) MS Access (.mdb/.accdb)
Quantitative tabular data with minimal metadata **a matrix of data with or without column headings or variable names, but no other metadata or labelling**	comma-separated values (CSV) file (.csv) tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate	delimited text of given character set - only characters not present in the data should be used as delimiters (.txt) widely-used formats, e.g. MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf) and OpenDocument Spreadsheet (.ods)
Geospatial data **vector and raster data**	ESRI Shapefile (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data	ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mif) for vector data Keyhole Mark-up Language (KML) (.kml) Adobe Illustrator (.ai), CAD data (.dxf or .svg) binary formats of GIS and CAD packages
Qualitative data **textual**	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml) Rich Text Format (.rtf) plain text data, ASCII (.txt)	Hypertext Mark-up Language (HTML) (.html) widely-used proprietary formats, e.g. MS Word (.doc/.docx) some proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti

Data storage and preservation – file formats

Digital image data	**TIFF version 6 uncompressed (.tif)**	**JPEG (.jpeg, .jpg) but only if created in this format** **TIFF (other versions) (.tif, .tiff)** **Adobe Portable Document Format (PDF/A, PDF) (.pdf)** **standard applicable RAW image format (.raw)** **Photoshop files (.psd)**
Digital audio data	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3) but only if created in this format Audio Interchange File Format (AIFF) (.aif) Waveform Audio Format (WAV) (.wav)
Digital video data	MPEG-4 (.mp4) motion JPEG 2000 (.mj2)	
Documentation and scripts	Rich Text Format (.rtf) PDF/A or PDF (.pdf) HTML (.htm) OpenDocument Text (.odt)	plain text (.txt) some widely-used proprietary formats, e.g. MS Word (.doc/.docx) or MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0

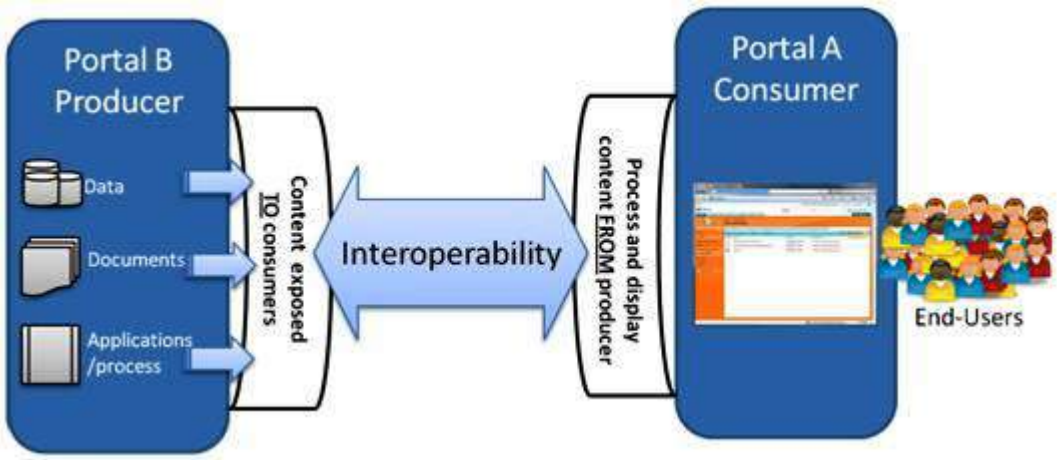
http://www.data-archive.ac.uk/create-manage/format/formats-table

3.5 Limited internet connectivity

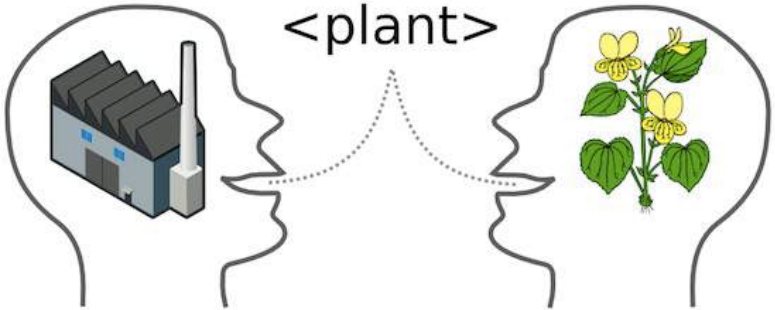
Internet connectivity: Access, affordability, ability

- Easily accessible information products (e.g. websites, PDFs)
- Alternative versions that require minimal data download
- Mobile versions
- Hosting/server choices based on connectivity – is cloud computing for everyone?

3.2 Interoperability



- Syntactic interoperability: communicate and exchange data
- Semantic interoperability: ascribe meaning to and automatically interpret data (via common vocabularies, trait dictionaries etc)



Consortium
CGIAR

CGIAR works across centers via 16 CRPs

- **MAIZE**
- **WHEAT**
- **GRiSP (Global Rice Science Partnership)**
- **Roots, Tubers & Bananas**
- **Dryland Cereals**
- **Grain Legumes**
- **Livestock & Fish**

- **CRP for Managing & Sustaining Crop Collections**

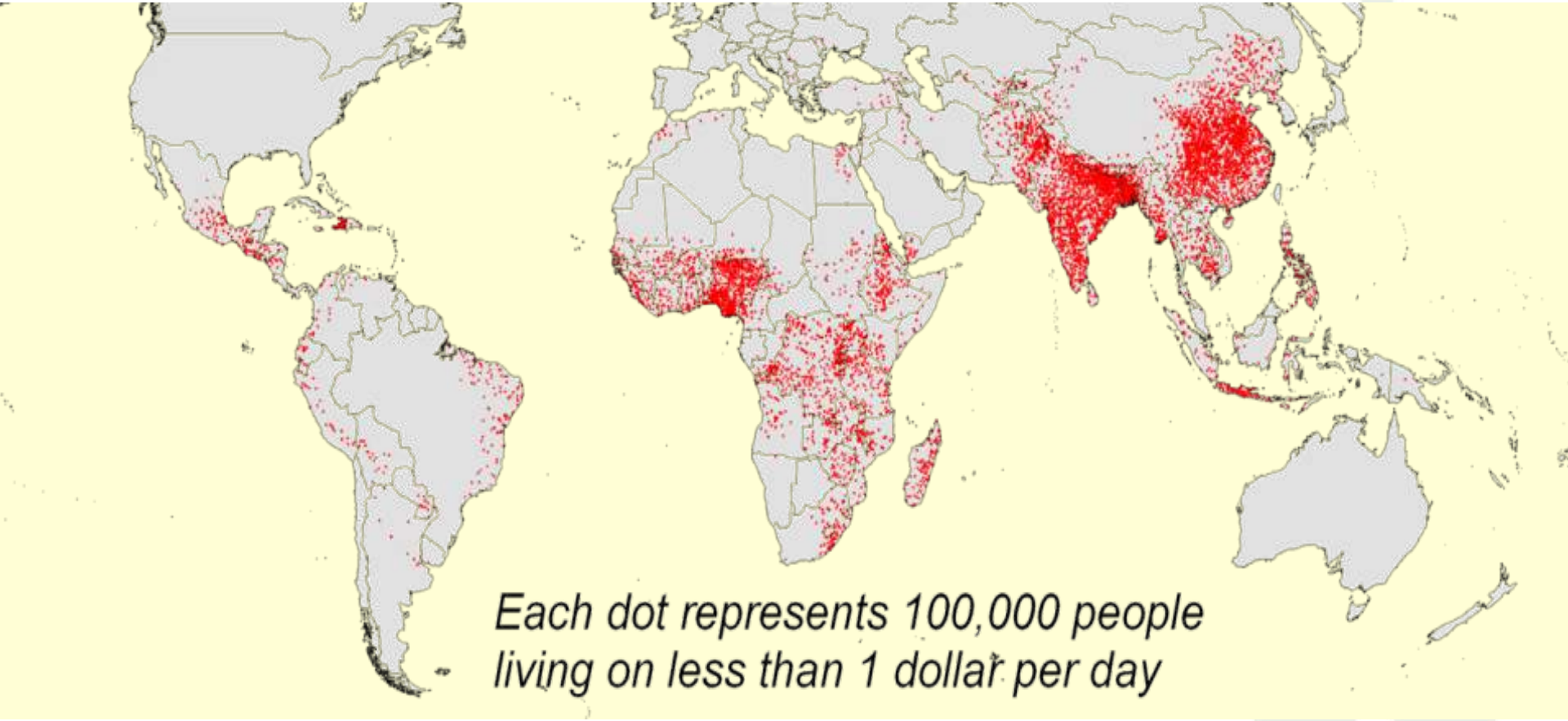
- **Policies, Institutions & Market**

- **Agriculture for Nutrition & Health**

- **Humid Tropics**
- **Aquatic Agricultural Systems**
- **Dryland Systems**

- **Climate Change, Agriculture and Food Security (CCAFS)**
- **Forests, Trees and Agroforestry (FTA)**
- **Water, Land and Ecosystems (WLE)**

Poverty: 1.2 billion on less than USD 1 per day



CGIAR repository systems - publications

- Almost all Centers have/are finalizing publication repositories

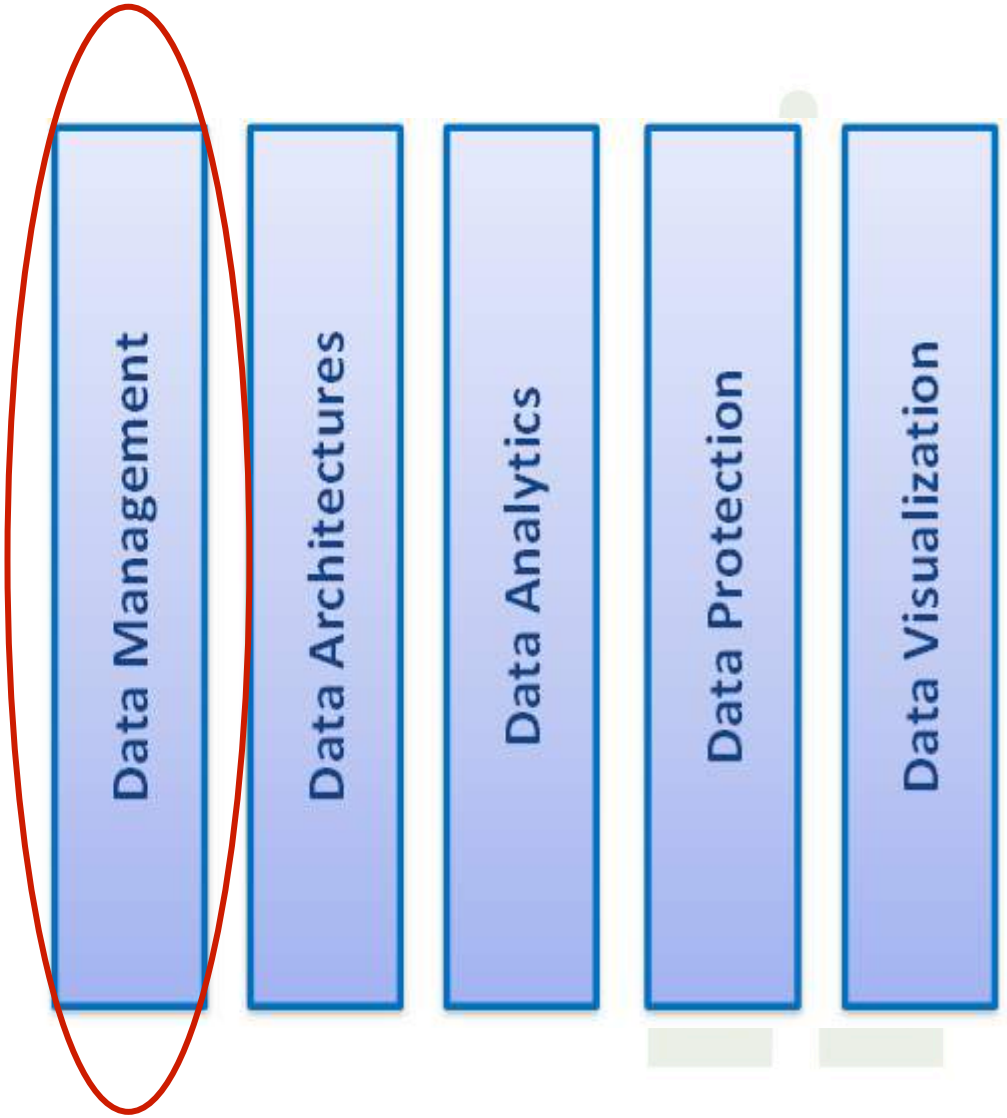
	Item-Level Access	Launch Year	# Items	Repository Platform
Africa Rice	Metadata only	2009	579	Other - Mendeley
Bioversity	Metadata only	2013	70	DSpace
CCAFS	Metadata & full text	2012	643	DSpace
CIAT	Metadata & full text	2012	6700	DSpace
CIFOR	Metadata only	2009	3000	Other - InMagic
CIMMYT	Full text	2010	3500	DSpace
CIP				(Dspace)
ICARDA	Full text	2011	500	Other – SharePoint 2013
ICRAF	Full text	2014	6240	Other - Invenio
ICRISAT	Metadata & full text	2009	7000	EPrints
IFPRI	Metadata & full text	2011	10000	OCLC/ContentDM
IITA	Full text	2009	8000	Other - Aigaion
ILRI	Metadata & full text	2009	12000	DSpace
IWMI	Full text		2390	DSpace
WLE	Metadata & full text	2013	450	DSpace

CGIAR repository systems - data

- Several Centers have data repositories – several developing…

	Data Repository	Repository Platform	Launch Year	# Items
Africa Rice	Yes	Dataverse	2012	24
Bioversity	Yes	Dataverse	2013	4
CCAFS	Yes	Dataverse	2012	108,103
CIAT	Yes	Dataverse	2013	6
CIFOR	Yes	Dataverse	2013	55
CIMMYT	Under development	Dataverse	2012	
CIP	Under development	Datavarse/ Biomart	2008	100,000
ICARDA	Under development		2009	
ICRAF	Yes	Dataverse	2013	209
ICRISAT	Yes	Dataverse	2013	450
IFPRI	Yes	Dataverse	2008	107
IITA	Under development	CKAN		
ILRI	Under development	CKAN	2014	
IWMI	Yes	Dataverse?	2009	3000
WLE	No			

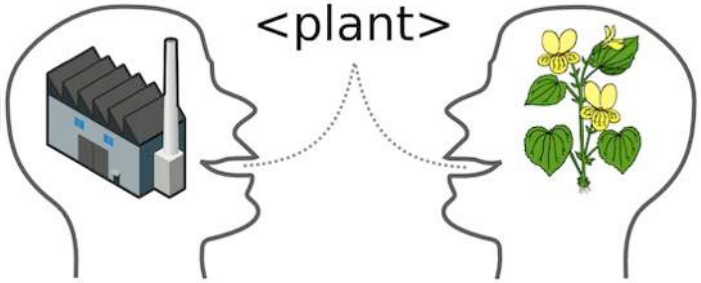
Getting “Big”…



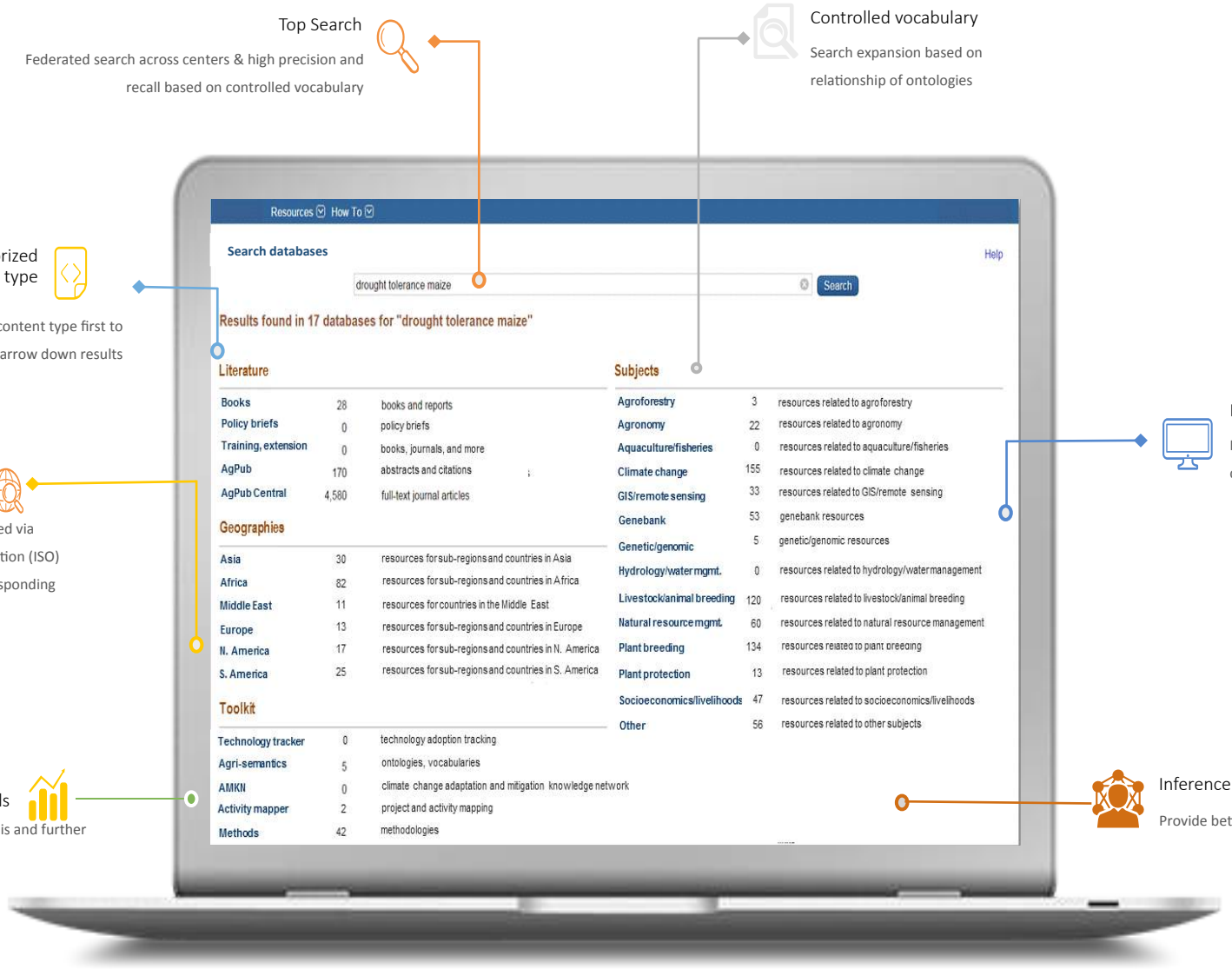
http://informationcatalyst.com/

Guiding principles for repositories

- Adherence to spirit of “openness” – CGIAR Policy; donor policies
- Repositories compliant with industry standards for semantic interoperability
  - OAI-PMH (metadata)
  - XML, RDF, JSON, O-data…



- Common minimum metadata schema → CG core
  - CG core application profile – v. Sep 2015
  - CG core – AgTrials mapping


Top Search
Federated search across centers & high precision and recall based on controlled vocabulary
Controlled vocabulary
Search expansion based on relationship of ontologies
Categorized content type
Browsed by content type first to narrow down results
Geo-referencing
Any contents referenced via standard geo-coordination (ISO) and easy to map corresponding region/countries
Machine-readable
Human and machine readable contents
Collection of tools
Toolkits for analysis and further researchers
Inference based on machine agents
Provide better customized search
Resources
How To
Search databases
Help
drought tolerance maize
Search
Results found in 17 databases for "drought tolerance maize"
Literature
Books 28 books and reports
Policy briefs 0 policy briefs
Training, extension 0 books, journals, and more
AgPub 170 abstracts and citations
AgPub Central 4,580 full-text journal articles
Geographies
Asia 30 resources for sub-regions and countries in Asia
Africa 82 resources for sub-regions and countries in Africa
Middle East 11 resources for countries in the Middle East
Europe 13 resources for sub-regions and countries in Europe
N. America 17 resources for sub-regions and countries in N. America
S. America 25 resources for sub-regions and countries in S. America
Toolkit
Technology tracker 0 technology adoption tracking
Agri-semantics 5 ontologies, vocabularies
AMKN 0 climate change adaptation and mitigation knowledge network
Activity mapper 2 project and activity mapping
Methods 42 methodologies
Subjects
Agroforestry 3 resources related to agroforestry
Agronomy 22 resources related to agronomy
Aquaculture/fisheries 0 resources related to aquaculture/fisheries
Climate change 155 resources related to climate change
GIS/remote sensing 33 resources related to GIS/remote sensing
Genebank 53 genebank resources
Genetic/genomic 5 genetic/genomic resources
Hydrology/water mgmt. 0 resources related to hydrology/water management
Livestock/animal breeding 120 resources related to livestock/animal breeding
Natural resource mgmt. 60 resources related to natural resource management
Plant breeding 134 resources related to plant breeding
Plant protection 13 resources related to plant protection
Socioeconomics/livelihoods 47 resources related to socioeconomics/livelihoods
Other 56 resources related to other subjects

Building a "cyberinfrastructure for agriculture"

- Infrastructure (Linked Open Data-enabled)
- Semantics + interoperability
- Toolkit (analysis, visualization, discovery, foresight, M+E…)
- Culture (advocacy, education, support to facilitate data-sharing)
- Partners (FAO, CABI, USDA, AgMIP, WUR, INRA, CIRAD…)