Pesquisa Orientada a Dados, Infraestrutura de Dados, e Desafios para a Ciência da Informação

Jian Qin
School of Information Studies
Syracuse University
Syracuse, New York, USA

XV ENANCIB, 27 de Outubro de 2014, Belo Horizonte, Brasil

Tradução para o portugues (*Translation to Portuguese*):
Kátia Cardoso Coelho (katiaccoelho@gmail.com)
Doutoranda do Programa de Pós Graduação em Ciência da Informação, UFMG

Orientação à Dados X

- Pesquisa orientada à Dados
- Política orientada à dados
- Tomada de decisão orientada à dados
- Negócio orientado à dados
- Orientado à dados …
- Orientado à dados …
- Orientado à dados …

School of Information Studies
SYRACUSE UNIVERSITY

Três temas desta apresentação

Pesquisa orientada à dados

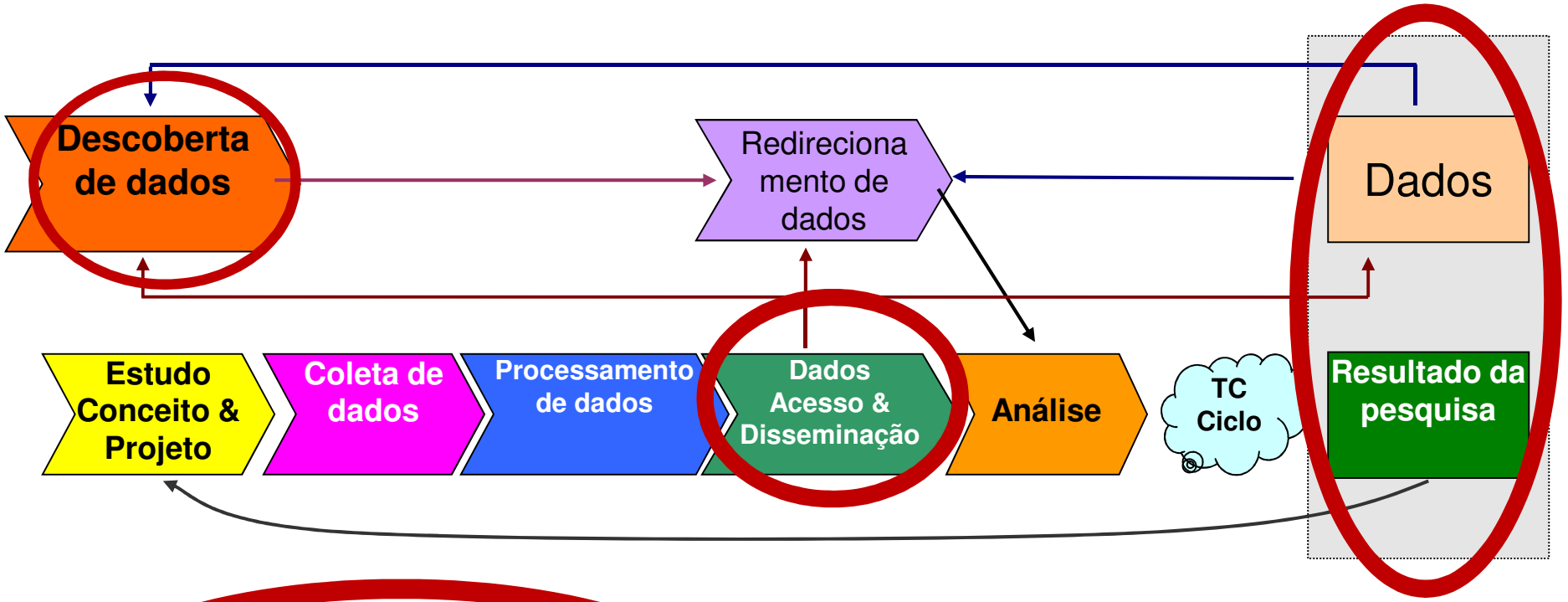
Novos territórios da biblioteconomia e serviços de informação

Desenvolver habilidades para serviço de dados

School of Information Studies
SYRACUSE UNIVERSITY

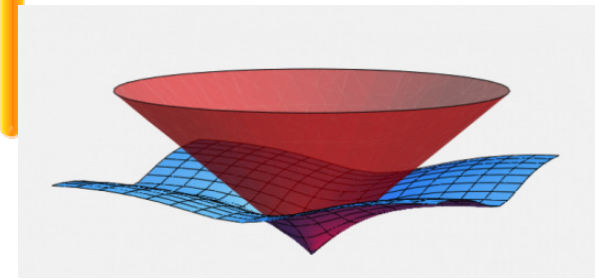
Pesquisa orientada à dados

E-Ciência e o ciclo de vida da pesquisa

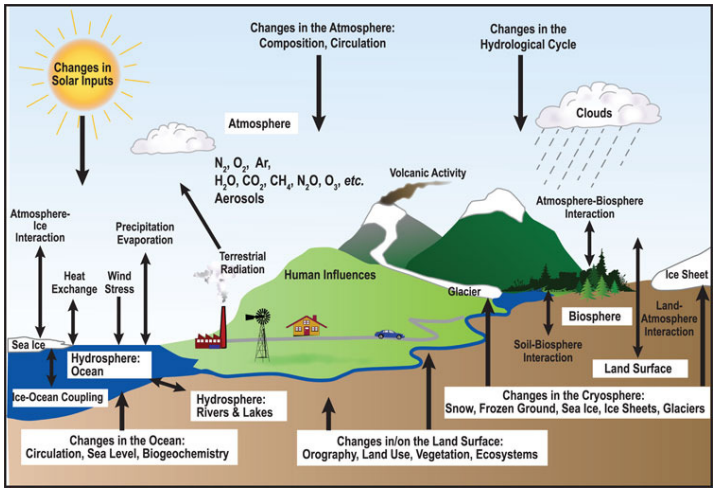


Onde os profissionais de informação e bibliotecários podem contribuir e proporcionar impactos

O "Ciclo TC" no diagrama representa o processo de Transferência de Conhecimento. Este diagrama de ciclo de vida vem de Charles Humphrey, "E-Science e Ciclo de Vida da Pesquisa" (2006) disponível em http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc



Mas o que é pesquisa?



School of Information Studies
SYRACUSE UNIVERSITY

Contexto de pesquisa

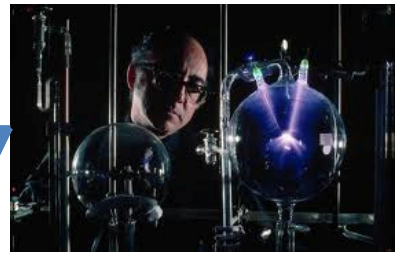
Acadêmico

Empresarial

Governo

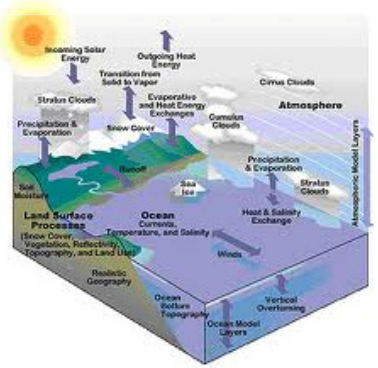
Sem
fins lucrativos

Tipos de pesquisa

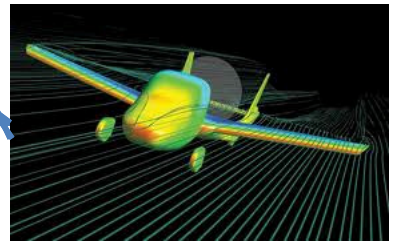


Experimental

Observação

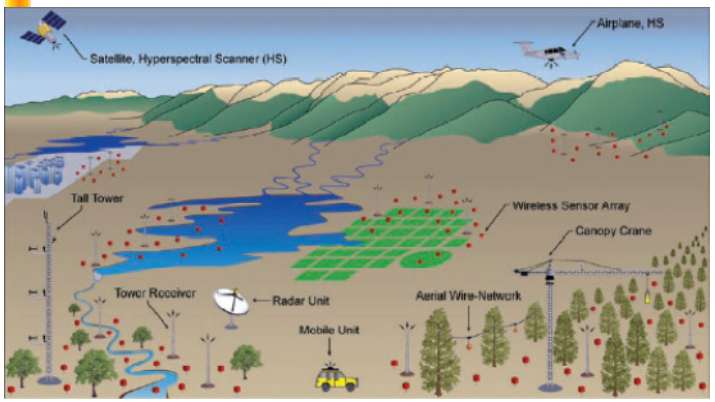


Modelagem



Simulação

Um cenário de coleta de dados



NSF. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Instrumentos que coletam dados

- Sensores
- Torres de micro-ondas
- Sensoriamento remoto

Primeiro nível de processamento

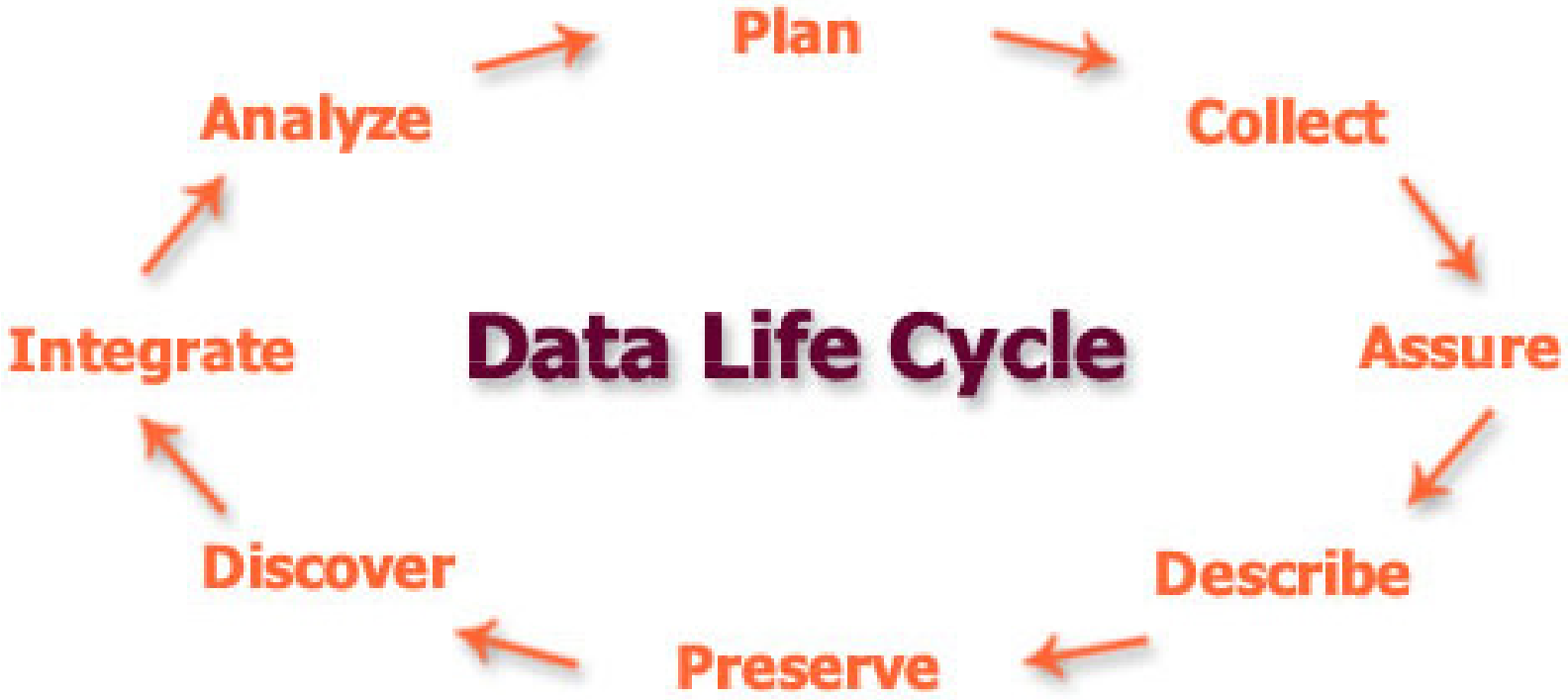
- Formatar
- Calibrar
- Documentar
- Arquivar (dados brutos)
- Entregar (cópias à equipe de pesquisa)

Nível dois de processamento:

- Organizar dados apropriadamente em arquivo de dados e segmentos
- Converter métricas / medições
- Devolver ao nivel dois de processamento cópias dos arquivos de dados

School of Information Studies SYRACUSE UNIVERSITY

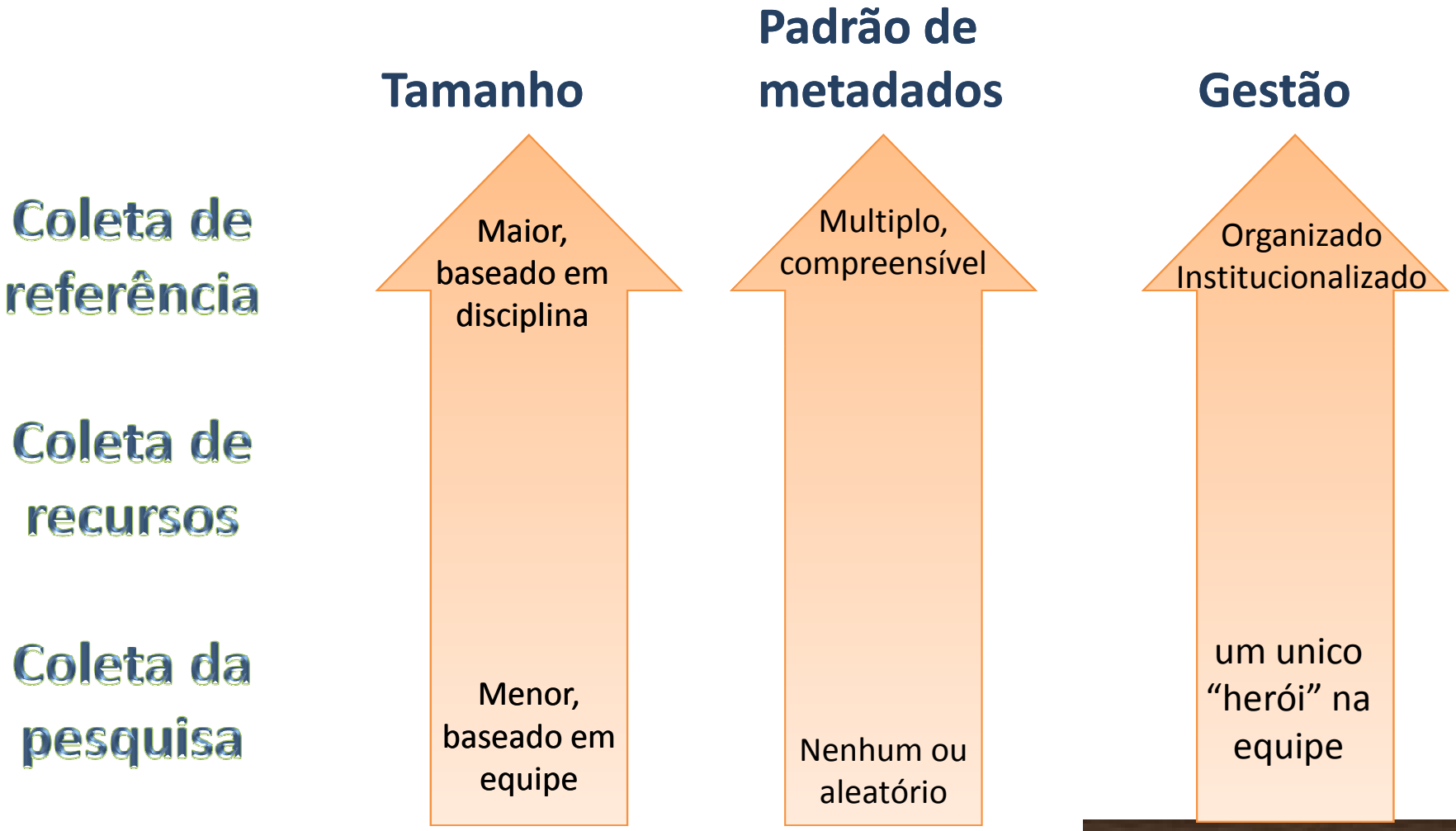
Ciclo de vida dos dados de pesquisa



http://www.dataone.org/best-practices

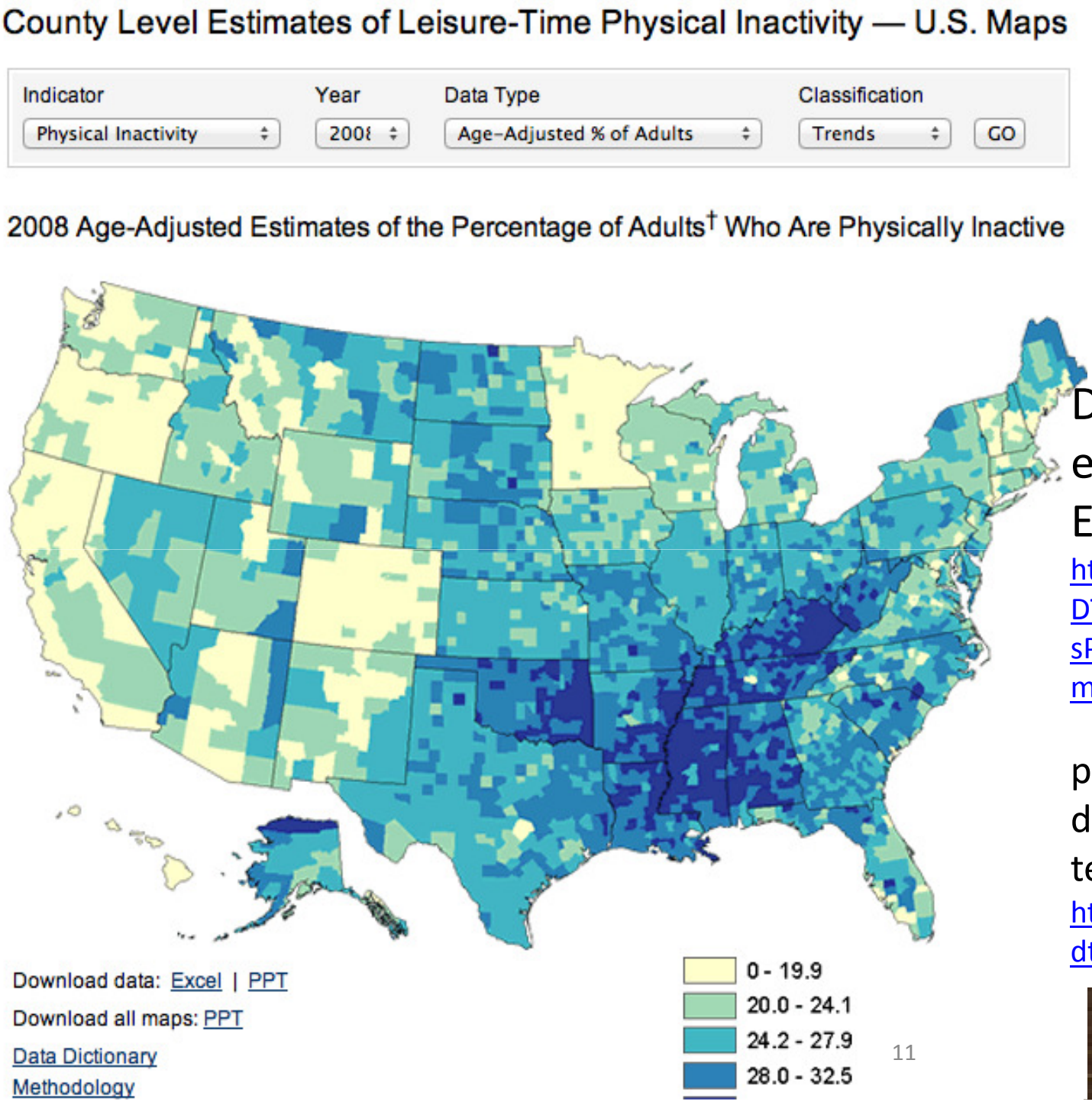
School of Information Studies
SYRACUSE UNIVERSITY

Coleta de dados de pesquisa



School of Information Studies SYRACUSE UNIVERSITY

Interativo dados produtos



Dados de diabetes e tendência — Estimativa do país: http://apps.nccd.cdc.gov/DDT_STRS2/NationalDiabetesPrevalenceEstimates.aspx?mode=PHY ;

pagina sobre dados de diabetes e tendências: http://apps.nccd.cdc.gov/ddtstrs/default.aspx

School of Information Studies
SYRACUSE UNIVERSITY

Registro de dados

Gestão de dados de ensaios clínicos: http://www.clinicaltrials.gov/ct2/show/NCT00006286?term=TADS+NIMH&rank=1

ClinicalTrials.gov
A service of the U.S. National Institutes of Health

Home Search Study

Study 1 of 1 for search of: TADS NIMH

Previous Study | Return to Search Results | Next Study

Full Text View | Tabular View | No Study Results Posted | Related Studies

Treatment for Adolescents With Depression Study (TADS)

This study has been completed.

First Received on September 14, 2000. Last Updated on January 18, 2008 History of Changes

Sponsor:	National Institute of Mental Health (NIMH)
Information provided by:	National Institute of Mental Health (NIMH)
ClinicalTrials.gov Identifier:	NCT00006286

Purpose

TADS is designed to compare the effectiveness of established treatments for teenagers suffering from major depressive disorder (MDD). The treatments are: psychotherapy ("talking therapy"); medication; and the combination of psychotherapy and medication. Altogether, 432 teenagers (both males and females) ages 12 to 17, will take part in this study at 12 sites in the United States.

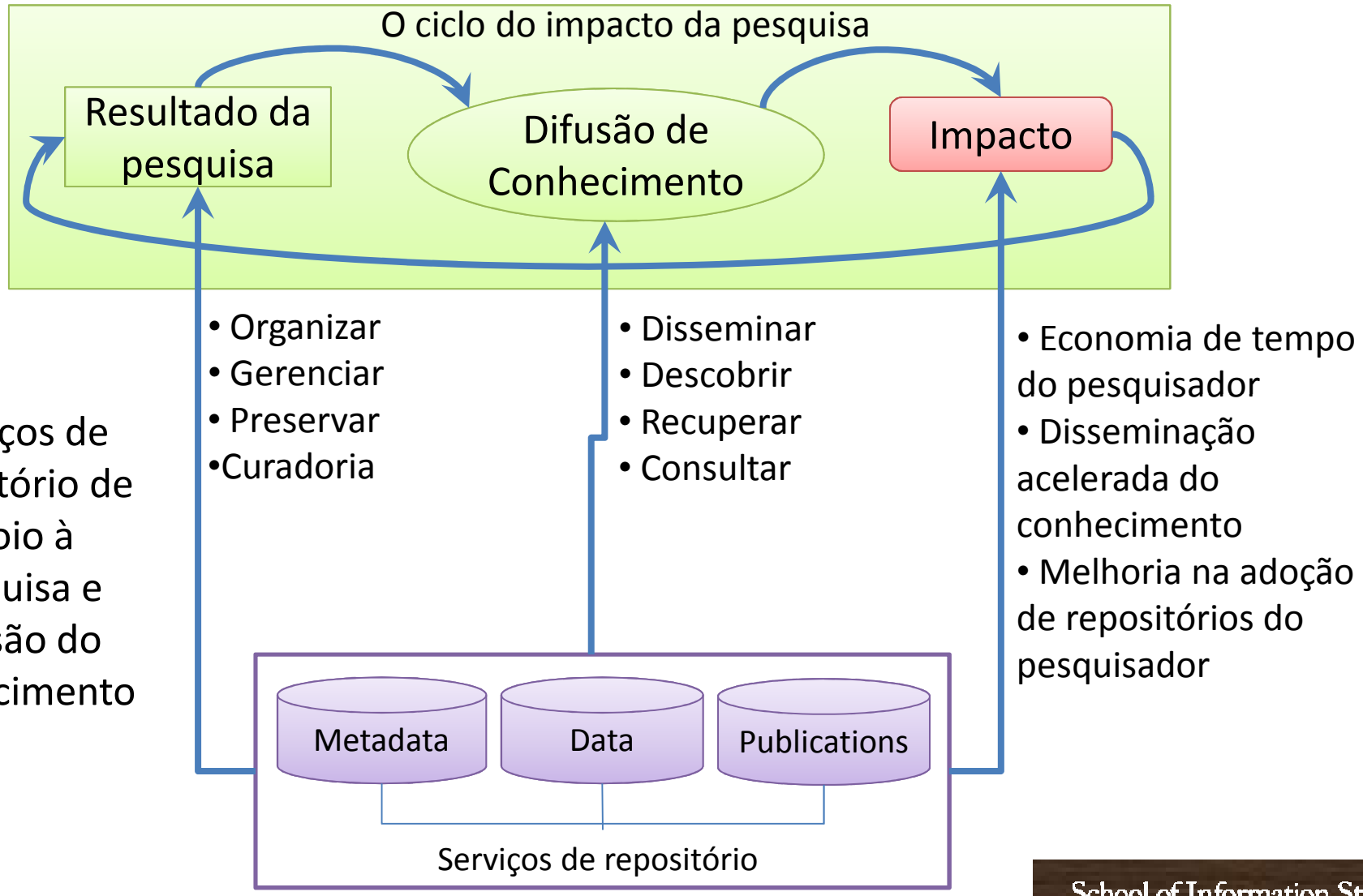
The **TADS** design will provide answers to the following questions: What is the long-term effectiveness of medication treatment of teenagers who have major depression? What is the long-term effectiveness of a specific psychotherapy ("talking therapy) in the treatment of teenagers who have major depression? How does medication treatment compare with psychotherapy in terms of effectiveness, tolerability and teenager and family acceptance? And, What is the cost-effectiveness of medication, psychotherapy and combined treatments?

The medication being used in this study is called fluoxetine. Fluoxetine is also known as Prozac. Research has shown that medications like Prozac help depression in young persons. Fluoxetine has been approved by the FDA for use in the treatment of child and adolescent (ages 7 to 17 years) depression.

The psychotherapy or "talking therapy" being used in this study is called Cognitive Behavioral Therapy (CBT). CBT is a talking therapy that will teach both the teenager and his or her family member (e.g., parent) new skills to cope better with depression. Specific topics include education about depression and the causes of depression, setting goals, monitoring mood, increasing pleasant activities, social problem-solving, correcting negative thinking, negotiation, compromise and assertiveness. CBT sessions may also help with resolving disagreements as they affect families.

School of Information Studies SYRACUSE UNIVERSITY

RDM, impacto da pesquisa, e valor



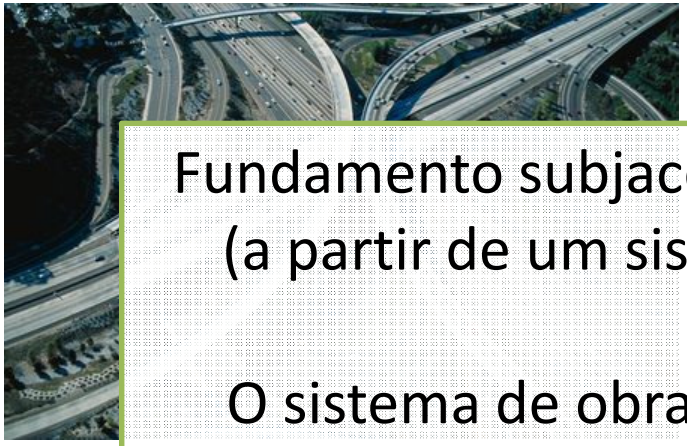
School of Information Studies SYRACUSE UNIVERSITY

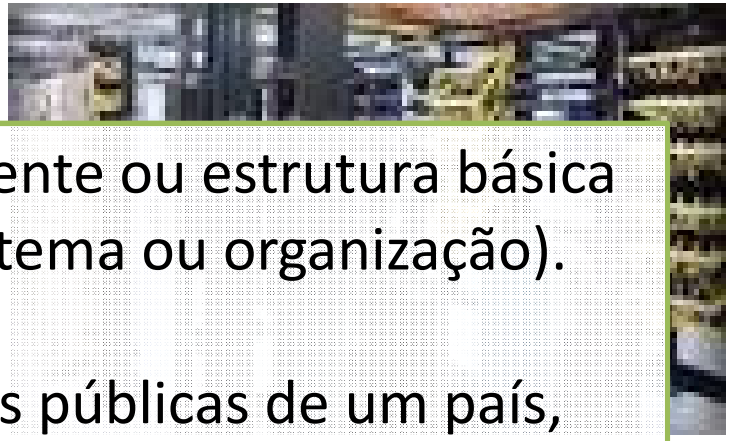
Resumo

- Cada estágio do ciclo de vida da pesquisa envolve algumas questões:
  - Baseada em ciência
  - Gestão de dados
  - Política
  - Técnica
- Objetivo da ciência de gestão de dados
  - Acesso à curto e longo prazo
  - Uso e reuso para vários propósitos por diferentes grupos de usuários

Pesquisa em gestão de dados como serviço de infraetrutura

O que é uma infrastrutura?

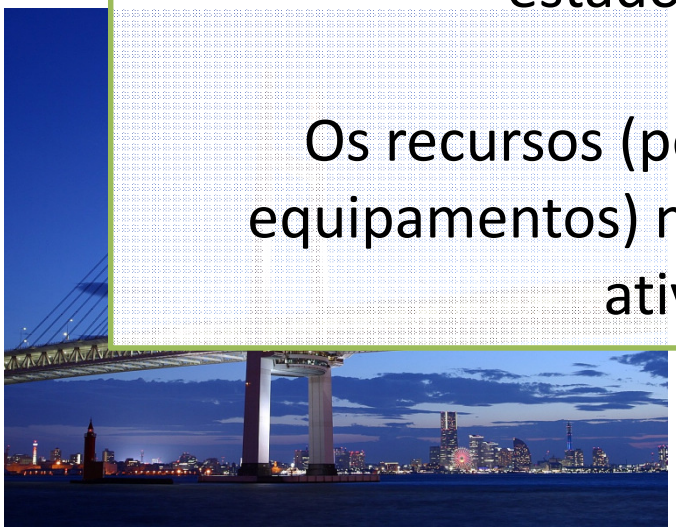


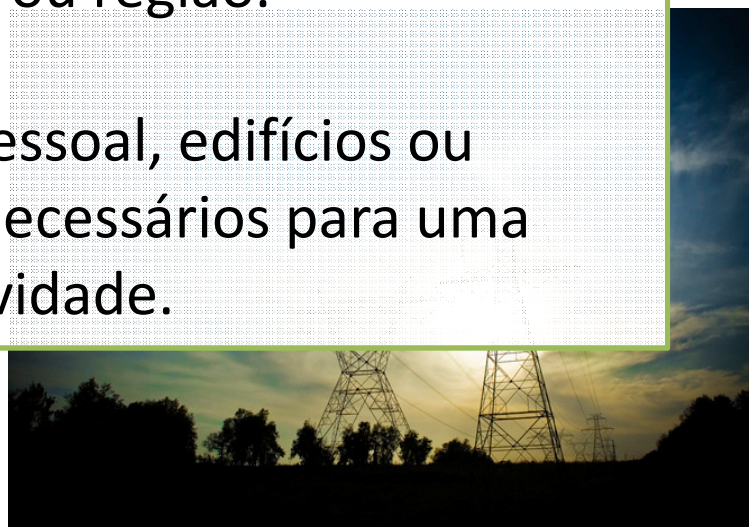


Fundamento subjacente ou estrutura básica (a partir de um sistema ou organização).

O sistema de obras públicas de um país, estado ou região.

Os recursos (pessoal, edifícios ou equipamentos) necessários para uma atividade.

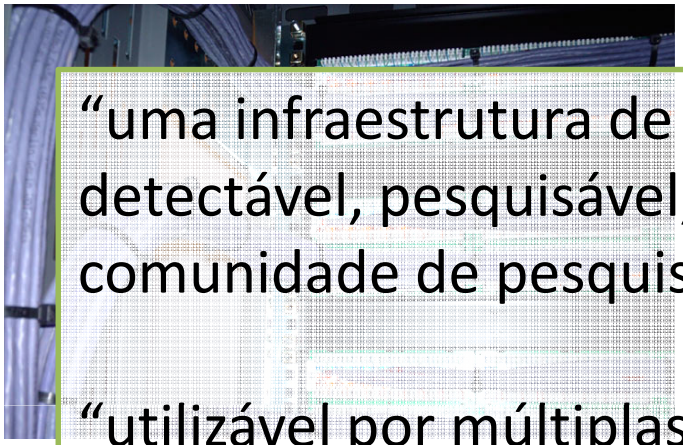




http://www.merriam-webster.com/dictionary/infrastructure

School of Information Studies
SYRACUSE UNIVERSITY

O que é uma infraestrutura de dados?

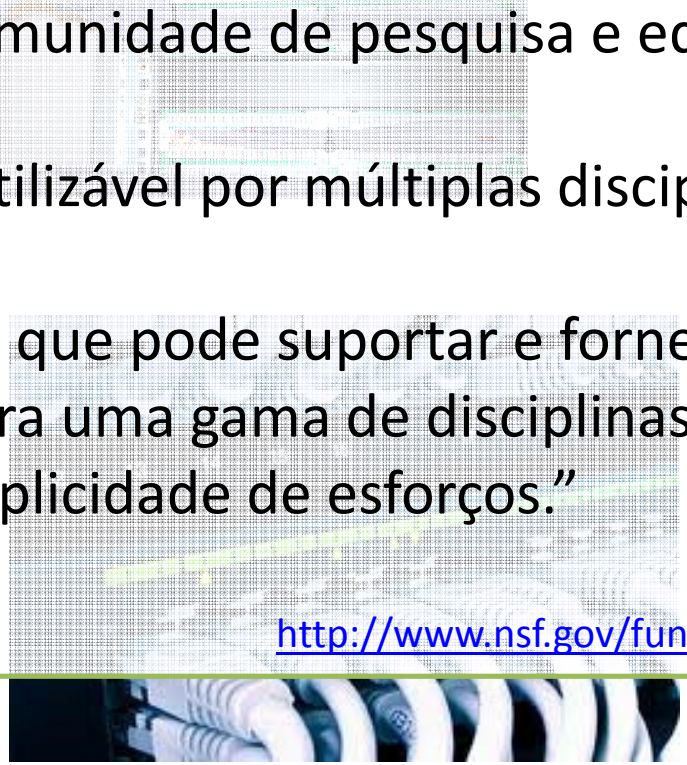


"uma infraestrutura de dados sustentável que será detectável, pesquisável, acessível e utilizável por toda comunidade de pesquisa e educação."

"utilizável por múltiplas disciplinas científicas..."

"... que pode suportar e fornecer soluções de dados para uma gama de disciplinas científicas, reduzindo a duplicidade de esforços."

http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504776



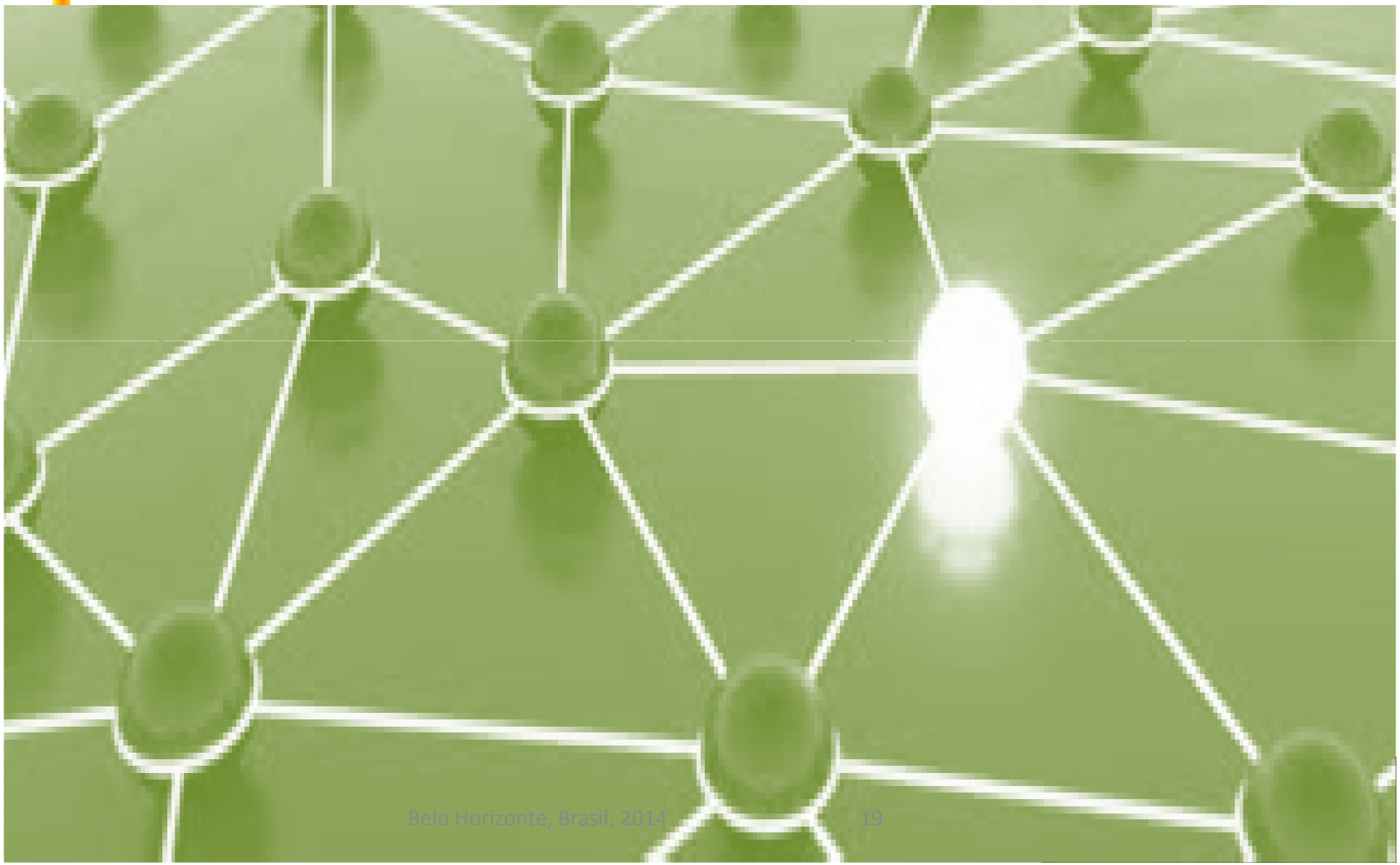
School of Information Studies
SYRACUSE UNIVERSITY

Natureza de uma infraestrutura

- **Enraizada.** Infraestrutura dentro de outras estruturas, arranjos sociais e tecnologias.
- **Transparente**. Infraestrutura não tem que ser reinventada para cada tarefa, mas suportar tarefas sem ser notada pelo usuário.
- **Alcance.** para além de um único evento ou uma prática local.
- **Aprendida como parte dos membros**.
- **Conectada com as convenções da prática**.
- **Incorporada a partir de padrões**.
- **Construída sobre uma base instalada**.
- **Visível ao se quebrar.**
- **Fixada em incrementos modulares, nem todos de uma vez ou globalmente.**

(Star & Ruhleder, 1996)

Conceitos importantes (1)



Conceitos importantes (2)

Ciber-infrastrutura: consiste de sistemas, sistemas de armazenamento de dados, instrumentos avançados e repositórios de dados, ambientes de visualização, bem como especialistas da computação, todos ligados por redes de alta velocidade para viabilizar a inovação acadêmica e descobertas.

Fonte: http://kb.iu.edu/data/auhf.html

Conceitos relevantes (3)

Coleções: itens de informação reunidos para um propósito específico ou com pelo menos uma característica em comum. Eles podem ser:

- gerados por uma instituição ou projeto
- reunidos para uma disciplina ou por um indivíduo

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf

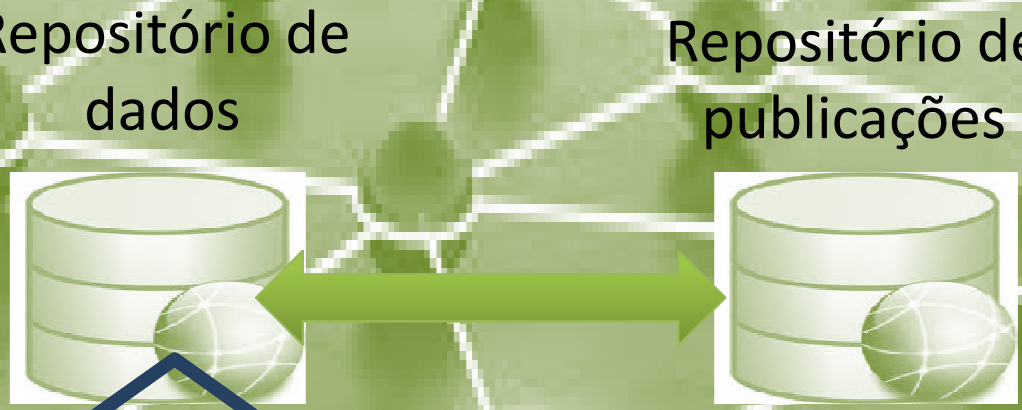
Repositórios: constructos que mantem coleções e facilitam seu uso.

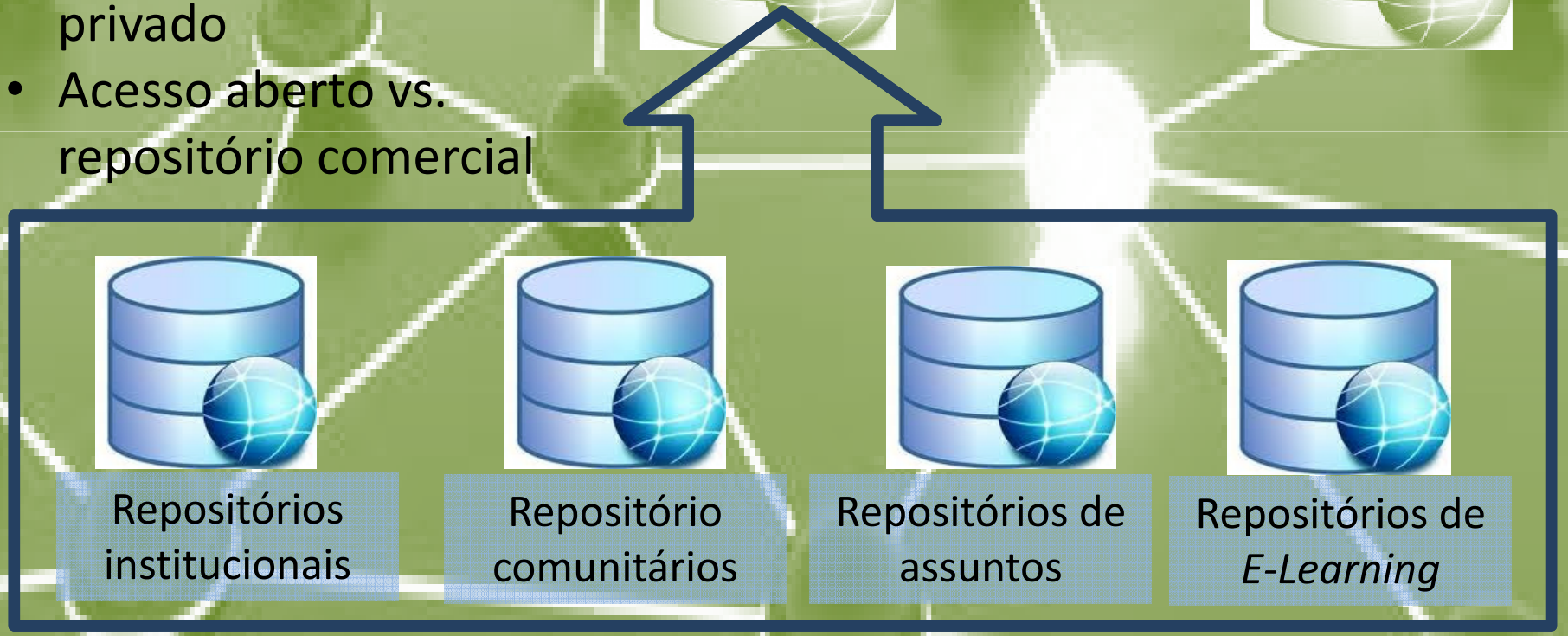
- Definição restrita: equipamentos de segurança e programas informatizados de apoio
- Definição ampla: incluem o quadro de gestão, serviços e ferramentas associadas a um repositório, bem como o próprio equipamento de armazenamento .

Conceitos importantes (3, cont...)

Outros conceitos relevantes relacionados a repositórios:

- Repositóro público vs. privado
- Acesso aberto vs. repositório comercial





Repositório como parte importante dos serviços de infraestrutura de dados

- Conteúdo é depositado no repositório por criador de conteúdo, proprietário ou terceiros
- A arquitetura do repositório gerencia os conteúdos e os metadados
- O repositório oferece um conjunto mínimo de serviços básicos, como inserir, extrair, pesquisar e controlar acesso
- O repositório deve ser sustentável, confiável, bem suportado e gerido

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf

Exemplos de serviços de infraestrutura de dados (abertos)

- *The Institute for Quantitative Social Science repository*: http://www.iq.harvard.edu/
- *Inter-University Consortium for Political and Social Research* (ICPSR): http://www.icpsr.umich.edu/icpsrweb/landing.jsp
- *The Dryad Digital Repository*: http://datadryad.org/
- *Data Observation Network for Earth*: http://www.dataone.org/
- *Datalib*: http://databib.org/ (a registry/directory/catalog of research data repositories)
- *Registry of Research Data Repositories*: http://www.re3data.org/

School of Information Studies
SYRACUSE UNIVERSITY

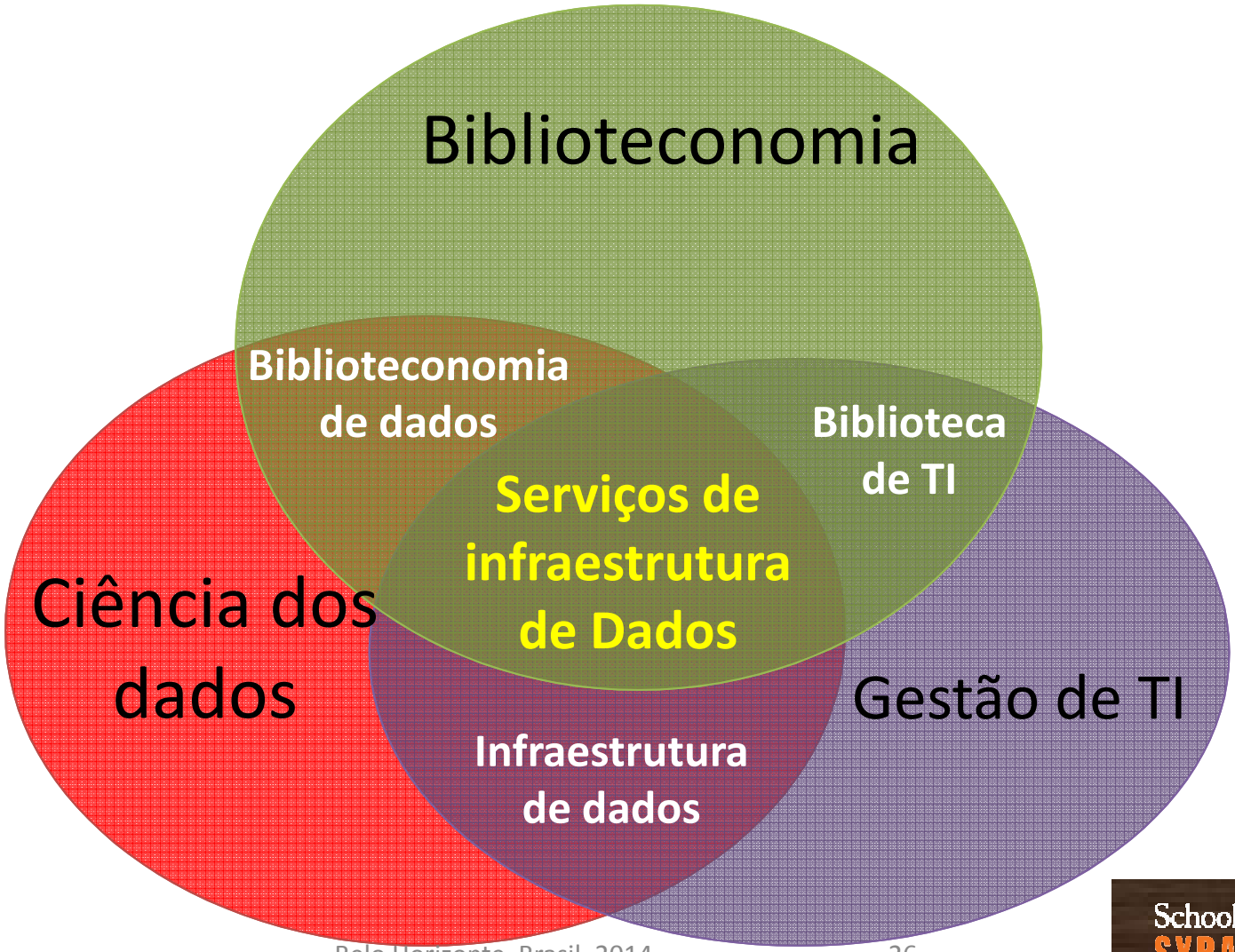
Ciência de dados: uma área emergente de trabalho

"Uma área emergente de trabalho relacionada com a coleta, apresentação, análise, visualização, gerenciamento e preservação de grandes coleções de informações."

(Stanton, 2012)

School of Information Studies
SYRACUSE UNIVERSITY

Serviços de infra-estrutura de dados e bibliotecas acadêmicas



School of Information Studies
SYRACUSE UNIVERSITY

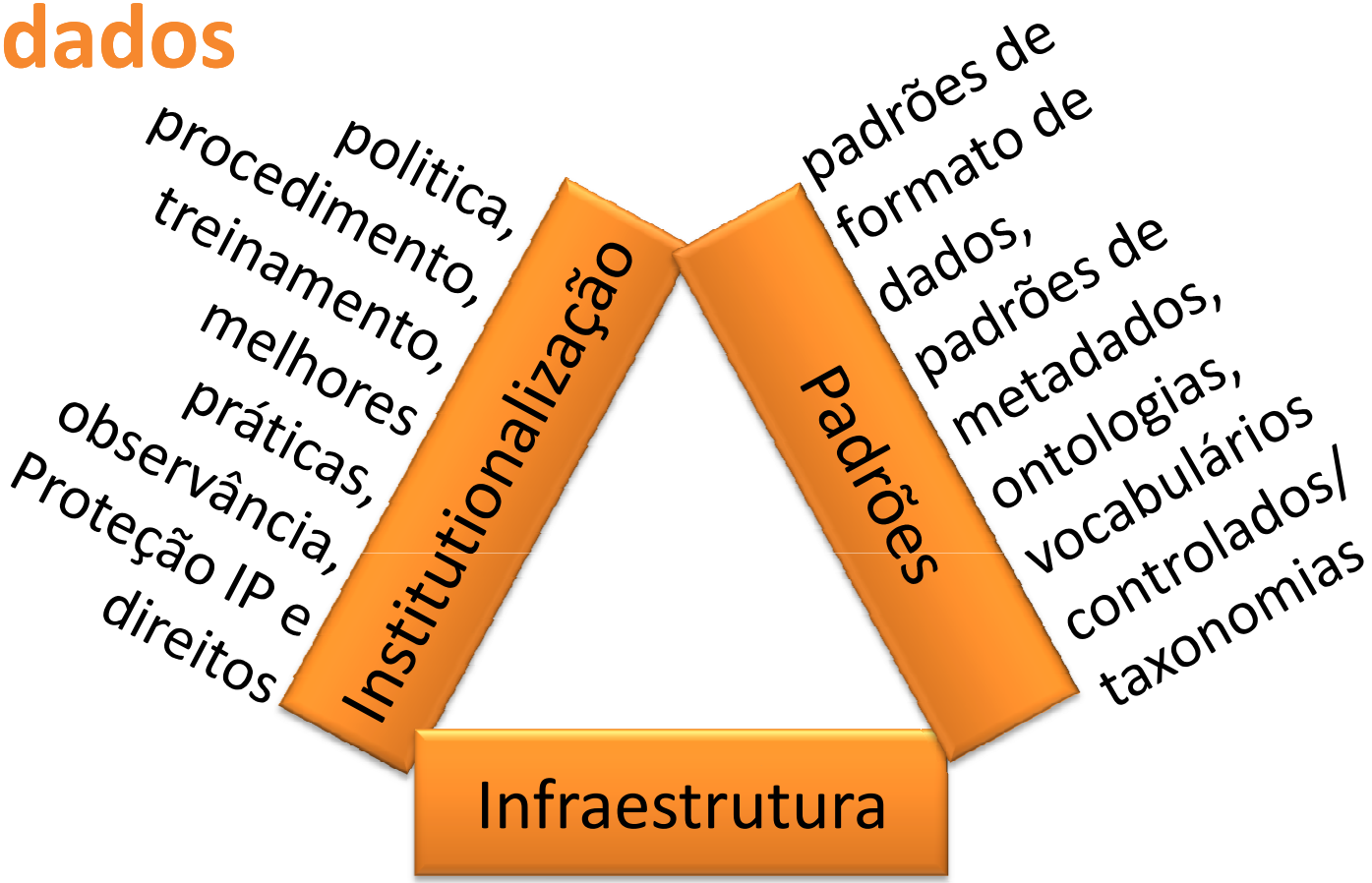
A palavra-chave para os serviços de infraestrutura de dados é:

Capacitação

Desafios e questões-chave no ambiente de pesquisa orientada à dados

(Áreas de investigação potencialmente importantes para a Ciência da Informação)

Três pilares para serviços de infraestrutura de dados



School of Information Studies SYRACUSE UNIVERSITY

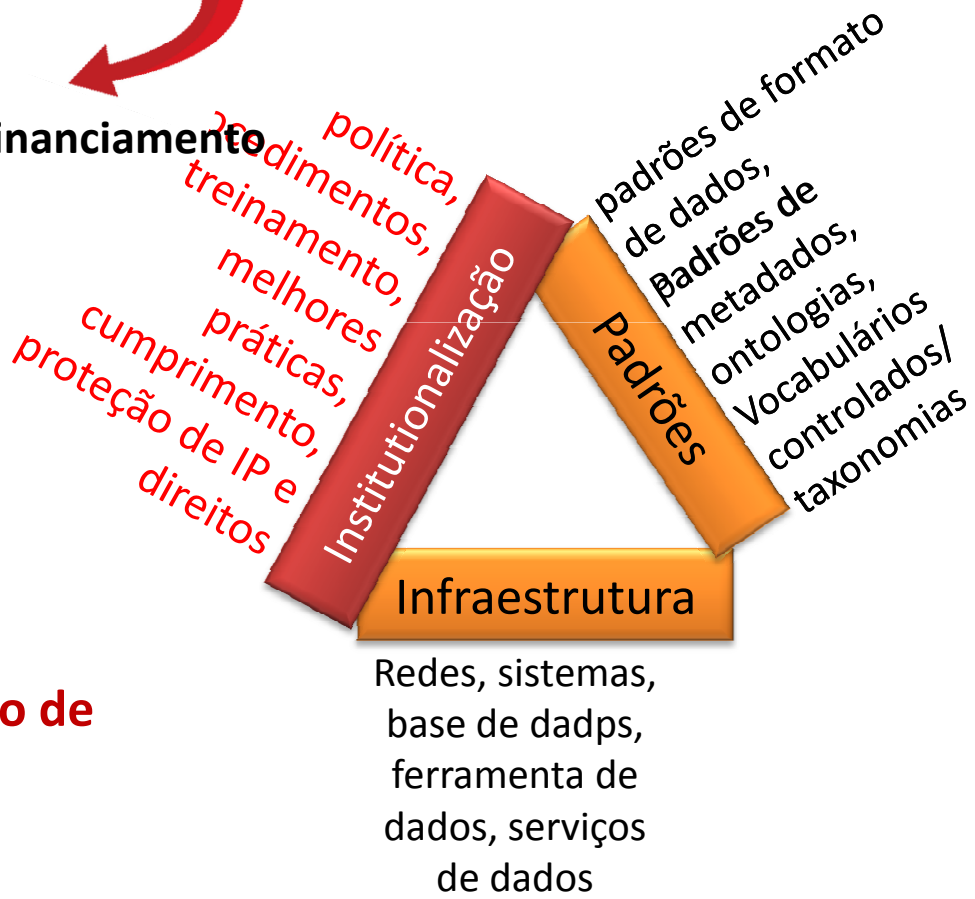
National Institutes of Health
Turning Discovery Into Health

National Science Foundation
WHERE DISCOVERIES BEGIN

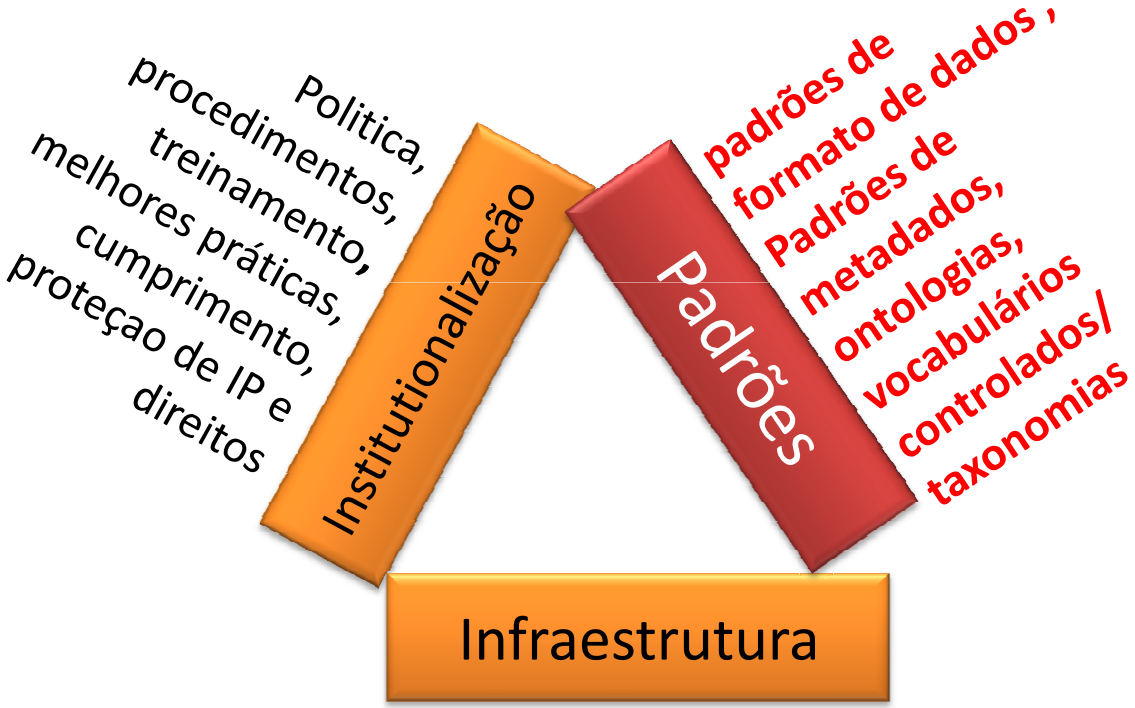
Mandatos políticos por agências de financiamento

Exigir que as instituições abordem estas questões :

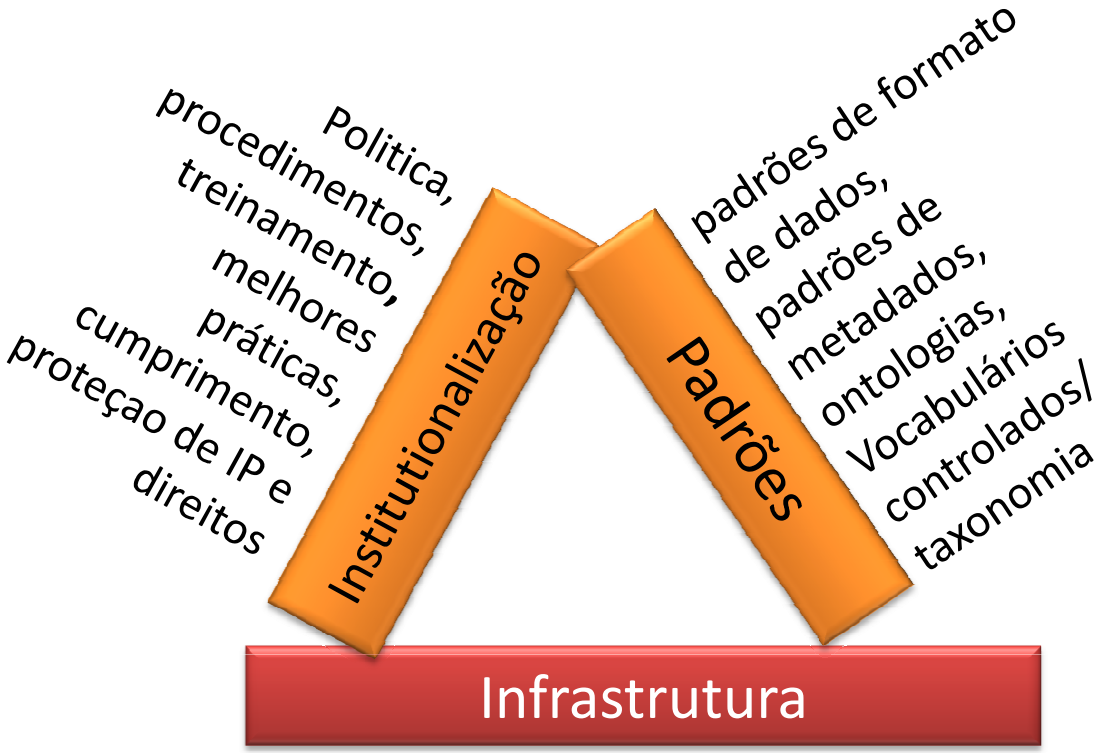
O que é institucionalização?
Por que institucionalizar a gestão de dados de pesquisa?
Como institucionalizar RDM?



Quanto se sabe sobre dados e metadados?
Como a natureza dos dados afeta os metadados?
Como os metadados afetam o acesso, compartilhamento, reutilização e preservação dos dados a longo prazo?



Rede, sistemas, base de dados, ferramenta de softwares, serviço de dados



Rede, sistemas, base de dados, ferramenta de softwares, serviço de dados

O que são infraestrutura de dados e serviços de infraestrutura de dados?

Por que construir uma infraestrutura de dados?

Qual é a chave para a construção de infraestruturas de dados?

O que está diante das bibliotecas de pesquisa?

Desafios

- Crescimento extremamente rápido e grande de volume de dados e metadados, dentro e fora dos repositórios
- Desenvolver e manter padrões, esquemas de banco de dados, ontologias etc, para melhorar a interoperabilidade
- Iniciativas começaram a criar infraestruturas de informação global ao nível semântico
- DIficuldades relativas ao depósito de dados, tanto técnica quanto comportal
- Repositórios são insuficientes

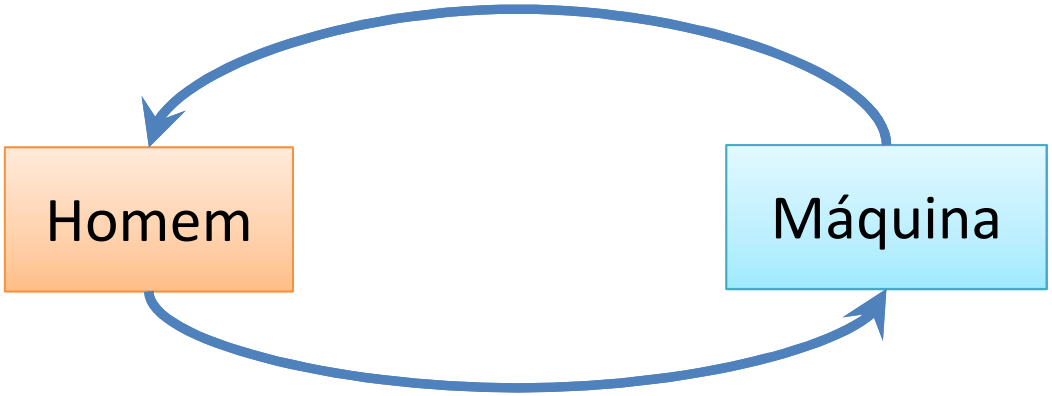
Onde estão os gaps?

Visão

- Uma infraestrutura de dados confiável e sustentável de serviços de repositórios de dados
- Um espaço de informação de alta qualidade
- Uma infraestrutura bem gerenciada de repositório responsável

Entendendo os serviços de dados

Para quem?



Tipo de infraestrutura de serviços:

- Nacional
- Institucional

School of Information Studies
SYRACUSE UNIVERSITY

Os serviços de dados têm evoluído…

ARL relatório de pesquisa 2010

Encontrar dados relevantes 83%

Desenvolver planos de gestão de dados 79%

Encontrar e usar a infraestrutura de tecnologia e ferramentas 76%

Desenvolver ferramentas para auxiliar pesquisadores 76%

Arquivar e prover curadoria de dados relevantes e curadoria para a preservação a longo prazo , além de integração

Prestação de serviços de curadorias e de administração de dados

Sensibilização e formação dos usuários

(Soehner, Steeves, & Ward, 2010)

ARL relatório de pesquisa 2013 (N=72)	
Fornecer um repositório institucional	89%
Localizar e usar fontes de dados existentes	94%
Apoiar GIS e análise geoespacial	85%
Comprar conjuntos de dados, aquisição e assinaturas	81%
Direitos autorais e patentes	75%
Suporte de software estatístico	58%
Suporte de visualização de dados	36%
Suporte a análise de dados	39%
Mineração de dados	28%
Projeto e gestão de banco de dados	28%
Programação / Desenvolvimento de software	24%
Outros serviços de apoio dados	29%

(Fearon et al., 2013)

Serviços do Plano de Gestão de Dados (PGD)

- Serviços Online PGD :
  - Explicação dos requisitos BPF por diferentes agências de financiamento e direções NSF
  - Diretrizes para a criação de PGDs
  - Exemplos de modelos de PGDs
  - Ferramenta ou recurso para criação de PGDs
  - Lista de verificação de planejamento de dados
  - Considerações sobre direitos autorais, diretrizes de citação de dados, exemplos de metadados, informações sobre os serviços de repositório digital
  - …

Quais os serviços de dados? (2)

- submissão de dados
- exportação de dados
- Formato de dados de conversão / transformação
- Acesso aos dados (descoberta e obtenção de dados)
- Gestão e Proteção IP
- Ofertas educativas
- Assistência técnica, incluindo gestão de dados e serviços de manipulação
  - Acesso às instalações
  - Curadoria
  - Ferramenta de arquivo e preservação
  - Informação
- Serviços de impressão e publicação
- *Marketing*
- Publicidade
- Serviços de desenvolvimento de Softwares

Fonte: Marcial & Hemminger, 2010

Questão 1: Recursos humanos

- ARL relatório de pesquisa (2013):
  - Bibliotecário de conteúdo ou bibliotecário de comunicação (50%)
  - Digital (38%)
  - Bibliotecário de dados (18%)
  - Metadados (17%)
  - Serviço de dados(13%)
  - GIS or Geospacial (12%)
  - Pesquisa em dados(11%)
  - Curadoria (11%)
  - Repositorio (10%)
  - Software ou sistemas (9%)
  - Gestão de dados (9%)

Questão 2: achados

- Orçamento normal da biblioteca interna 98%
- Financiamento da administração direta (separados dos fundos da biblioteca) 11%
- Concessão de financiamento externo 11%
- Orçamento temporário da biblioteca interna ou orçamento de projeto especial 9%
- Departamento ou instituto de pesquisa / projeto de fundos de grupo 6%
- Fundo de dotação 6%
- Taxa para concessão de pesquisador ou pesquisadora 4%
- Instalações e financiamento administrativo 2%
- Outra fonte de financiamento 9%

(N=57, from Fearon et al., 2013)

Questão 3: treinamento

- Treinamento/ Experiência mais importante para PGD:
  - *Expertise* de domínio Assunto
  - treinamento em curadoria digital/dados
  - Tecnologia de TI ou experiência em serviços
  - Treinamento e biblioteca MLS/MLIS

- Particularmente importante:
  - Métodos de pesquisa e análise de dados
  - Gerenciamento de dados de pesquisa
  - Curadoria de dados
  - Comunicação científica
  - Outras habilidades e treinamentos:
  - Identificação e aplicação de padrões de metadados
  - Preservação digital
  - Políticas de propriedade de dados
  - Questões éticas e legais
  - Segurança de dados
  - Compartilhamento de dados e acesso
  - Armazenamento e planejamento de backup de dados
  - Política de retenção de dados
  - Citação de dados

Questão 4: política de dados

- Políticas de dados institucionais
- Política de gestão de dados

Políticas de dados são um componente importante da institucionalização da gestão de dados de pesquisa

School of Information Studies SYRACUSE UNIVERSITY

Como você pode contribuir? (1)

- Capacidades do profissional da informação da biblioteconomia do século 21
  - Assunto complexo, processo ou expertise técnica
  - Compromisso de serviço complexo
  - Compromisso com Pesquisa e Desenvolvimento
  - Comprometimento para estimativa e avaliação
  - Habilidades em comunicação e *marketing*
  - Desenvolvimento de Projetos e Gestão de Competências
  - Engajamento político
  - Habilidades de desenvolvimento de recursos
  - Compromisso com o rigor
  - Espírito empreendedor
  - Compromisso com a colaboração
  - Liderança / Capacidade de Inspiração

http://www.arl.org/storage/documents/publications/2012-hrsym-pres-neal-j.pdf

Como você pode contribuir? (2)

"Funcionários da biblioteca acadêmica devem integrar os serviços de bibliotecas digitais, tradicionais, o arquivamento digital e de preservação, o desenvolvimento de repositório, a publicação digital e as tecnologias de ensino com o núcleo da biblioteca, do orçamento, de pessoal e da organização."

"Funcionários da biblioteca acadêmica devem estar incorporados nos processos de ciber-infraestrutura de pesquisa e comunicação científica, e ser parte integrante dos sistemas de gerenciamento de informações de pesquisa."

- Apoiar as necessidades de pesquisa orientada a dados
  - Agência Federal / financiamento
  - Grandes conjuntos de dados
  - Curadoria de dados não estruturados
  - Extração
  - Distribuição
  - Colaboração
  - Visualização
  - Simulação
  - Preservação

http://www.arl.org/storage/documents/publications/2012-hrsym-pres-neal-j.pdf

Considerações finais

- Gestão de dados e serviços são um território novo e requerem novo pensamento sobre os papéis da biblioteca de pesquisa
- Amplas oportunidades para bibliotecários contribuírem para a construção de novas capacidades
- Aprender lições de iniciativas anterioress
- Transformar a imagem da biblioteca requer transformar primeiro a tradição das bibliotecas

Obrigado!


enancib
2014
Além das "nuvens":
expandindo as fronteiras
da Ciência da Informação

Referências

- Fearon, D. Jr., Gunia, B., Pralle, B. E., Lake, S., & Sallans, A. L. (2013). Research Data Management Services. Washington, DC: Research Library Association.
- Star, S.L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information space. Information Systems Research, 7(1): 111-134.
- Marcial, L. H. & Hemminger, B. M. (2010). Scientific Data Repositories on the Web: An Initial Survey. *Journal of the American Society for Information Science and Technology*, 61(10): 2029-2048.
- McCormick, T. (2009). A Web services taxonomy: not all about the data. http://tjm.org/public/Web-Services-Taxonomy_McCormick_v1.1.pdf
- Venugopal, S., Buyya, R., & Ramamohanarao, K. (2006). A taxonomy of data grids for distributed data sharing, management, and processing. *ACM Computing Surveys*, 38(1): http://arxiv.org/pdf/cs.DC/0506034
- Soehner, C., Steeves, C., & Ward, J. (2010). E-Science and data support services: A study of ARL member institutions. http://www.arl.org/bm~doc/escience_report2010.pdf
- Stanton, J. (2012). Introduction to Data Science. http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf