

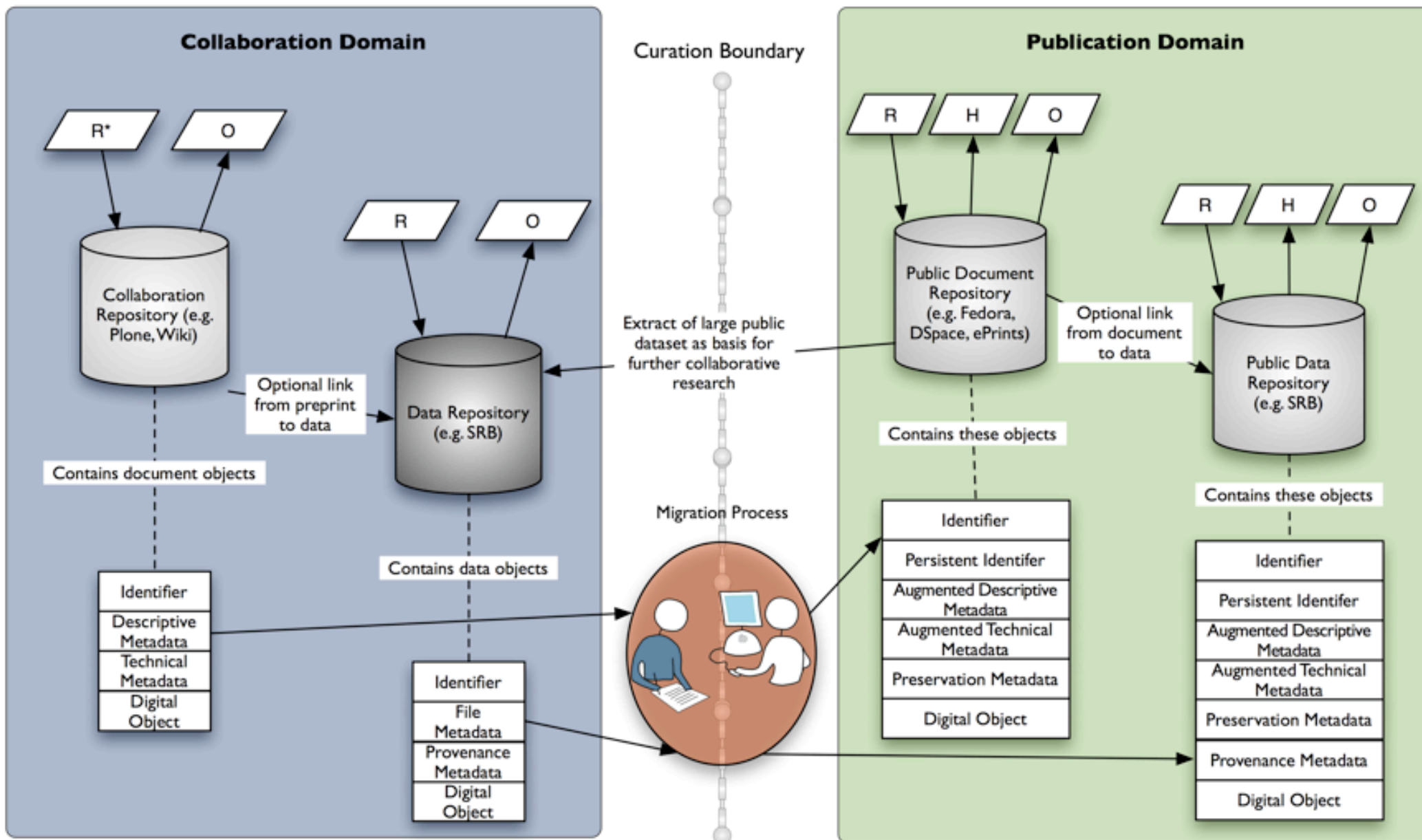
DONNÉES DE LA RECHERCHE IDENTIFIANTS



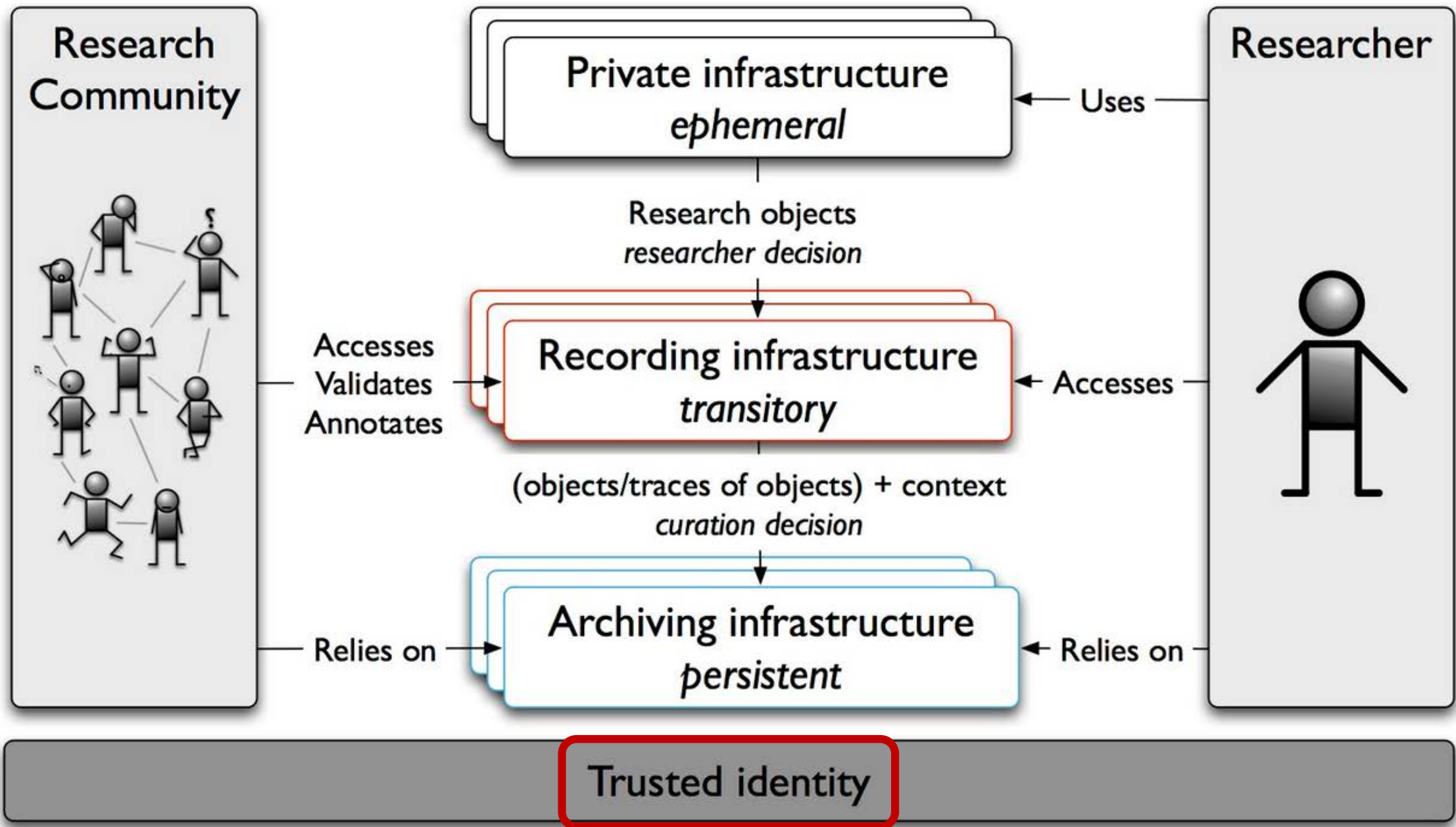
RENÉ SCHNEIDER
HAUTE ÉCOLE DE GESTION DE GENÈVE
LICENCE CC-BY 2.5

*All cartoons courtesy of Jørgen Stamp,
Digitalbevaring.dk. CC BY 2.5.*

Collaboration, Publication, and the Curation Boundary



* R = Register, H = Harvest, O = Obtain



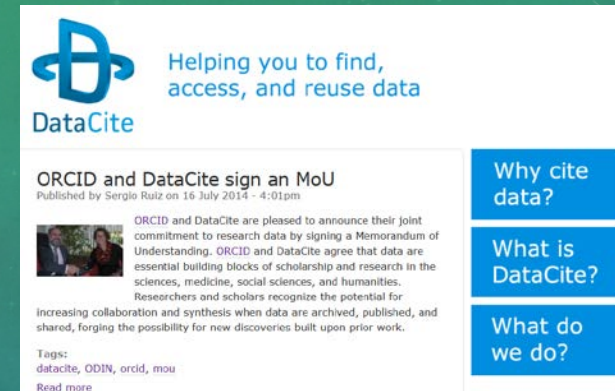
ENJEUX ET PROBLÈMES

- Définition: Qu'est-ce qu'un jeu de données, au juste?
- Granularité: Comment est-il structuré?
- Temps: A partir de quand a-t-on besoin d'identifiants
- Identification de l'auteur: Peut-il lui aussi être identifié de manière univoque?
- Micro-attribution: Quel chercheur est responsable de quelle étape dans le traitement des données?
- Comment traiter les données digitalisées après coup? Comment identifier les auteurs?

IDENTIFIANTS: UN APERÇU



DATA CITE & DOI



- DataCite est une organisation sans but lucratif, fondée pour l'enregistrement de jeux de données
- Les DOIs fonctionnent sur la base d'une combinaison d'identifiants pérennes, de métadonnées et d'un contrat social.
- Avantage: une bonne base pour l'identification pérenne de jeux de données, les données reçoivent un ID.
- Désavantage: aucune distinction n'est faite concernant la granularité des jeux de données ou des types de données, enregistrement pur, DataCite est au final un club payant.

CROSSREF



- Crossref a été fondé pour l'identification univoque de publications.
- Avantage: coopération avec DataCite
- Désavantage: voir DataCite, il s'agit au final dans les deux cas de silos de données centralisés

ARK, N2T



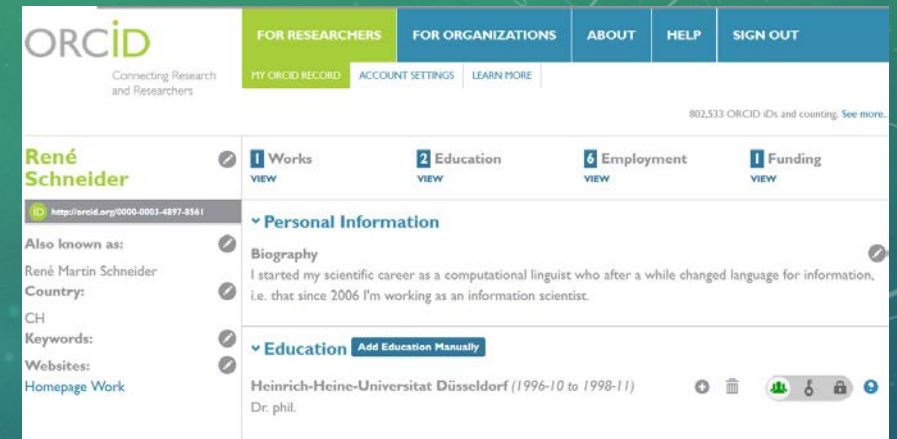
- N2T = Name to Thing Resolver
- Exemples: <http://soviet.org/ark:/12345/9876> = <http://russia.org/ark:/12345/9876>
- Une approche alternative et complémentaire pour l'archivage des données, qui ne s'intéresse pas aux adresses: décentralisé et agnostique des ID
- Les ARKs apparaissent à côté des DOIs dans le Thomson Reuter Data Citation Index
- Avantage: une alternative aux DOIs, suit l'approche classique d'archivage de contenus web, peut être géré localement, gratuit!
- Désavantage: difficile à comprendre, plutôt méconnus jusqu'à présent

EPIC



- EPIC attribue des identifiants dynamiques pour toutes les données émanant du processus de recherche, mais pas pour celles qui seront archivées.
- Intéressant pour les sciences devant comparer des sets de données non définitifs et à versions multiples.
- Avantage: complète l'offre d'identifiants
- Désavantage: un consortium, donc coûteux par nature

ORCID



- ORCID attribue des identifiants univoque aux auteurs des données
- Intégration: Lien entre le chercheur et son activité – affiliation
- Hub de relations entre différents identifiants, lisibles par machines
- Avantage: méthode prometteuse pour des identifiants de personnes
- Désavantage: uniquement pour auteurs vivants, car chaque auteur doit se créer son propre compte

ISNI



- ISNI attribue des identités aux vivants comme aux morts
- Indépendant des domaines, disciplines et rôles
- Pour une gestion des licences unique au sein de la recherche, des bibliothèques, de l'industrie des médias, des éditeurs et des autres institutions culturelles
- Avantage: génère des IDs pour les personnes mortes et les personnes fictives

DDI: DATA DOCUMENTATION INITIATIVE



- Initiative pour une documentation homogène des jeux de données
- Avantage: approche Linked Data
- Désavantage: encore trop peu connu



TRUSTED IDENTITIES

Des identifiants pérennes pour jeux de données ne signifient pas une attribution arbitraire de DOIs à n'importe quel ensemble de données.

Il en faut plus pour des «trusted identities».

Pour ce faire, plusieurs aspects doivent être considérés, en lien avec les dépôts, la dimension temporelle, la granularité des jeux de données, la paternité des données, etc.