AURA

Automatic Strengthening of Graph-Structured Knowledge Bases

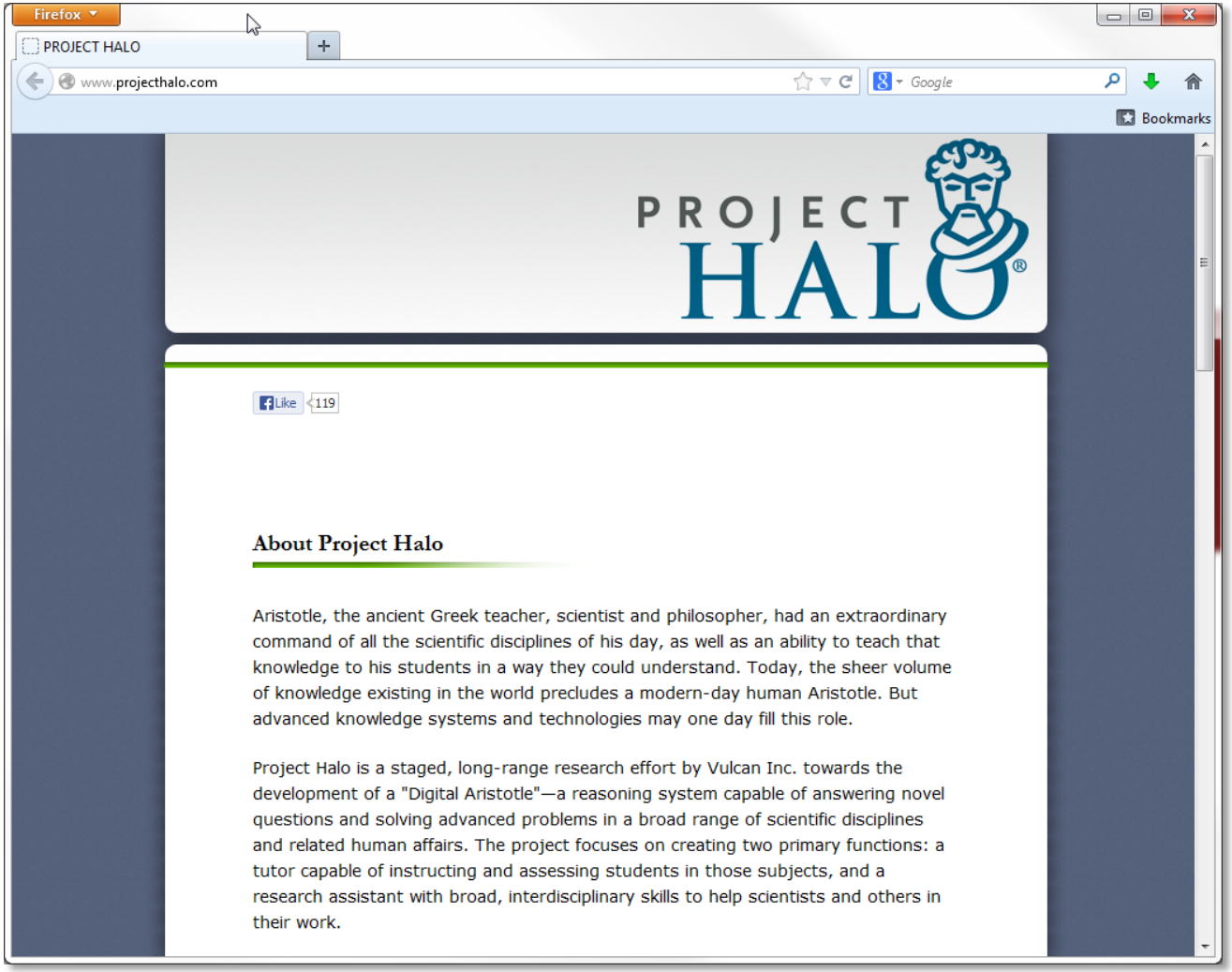
Vinay K. Chaudhri
Nikhil Dinesh
Stijn Heymans
Michael A. Wessel

SRI International

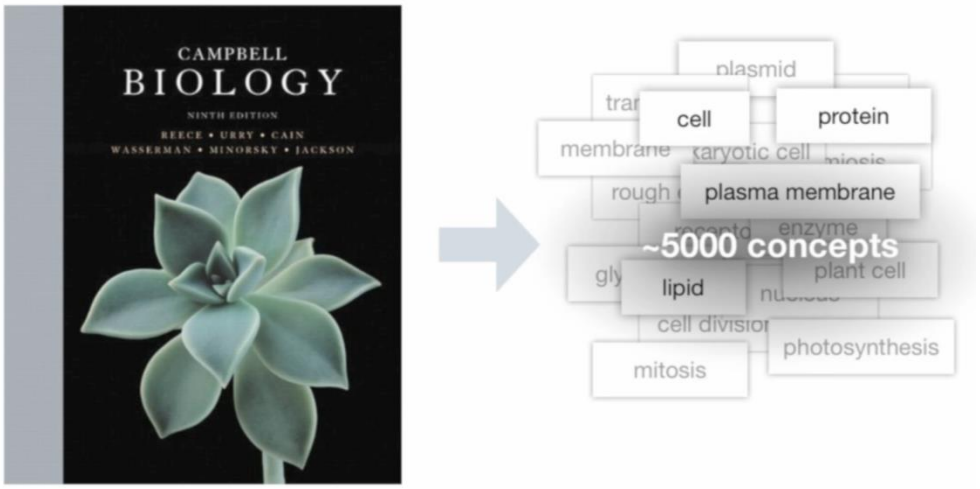
PROJECT HALO

Acknowledgment

- This work has been funded by Paul Allen's Vulcan Inc.
  http://www.vulcan.com
  http://www.projecthalo.com



The Biology KB of the AURA Project



- A team of **biologists** is using graphical editors to curate the KB from a popular Biology textbook, using a **sophisticated knowledge authoring process** (see http://dl.acm.org/citation.cfm?id=1999714)

- The KB is used as the basis of a smart question answering text book called **Inquire Biology** – questions are answered by AURA using forms of deductive reasoning

- The KB has non-trivial graph structure and is big (5662 concepts)

- The KB is a valuable asset: it contains 11.5 man years of biologists, and estimated 5 (2 Univ. Texas + 3 SRI) years for the upper ontology (CLib)

Graphical Modeling in AURA

S1 Every *Cell* has part a *Ribosome* and a *Chromosome*.

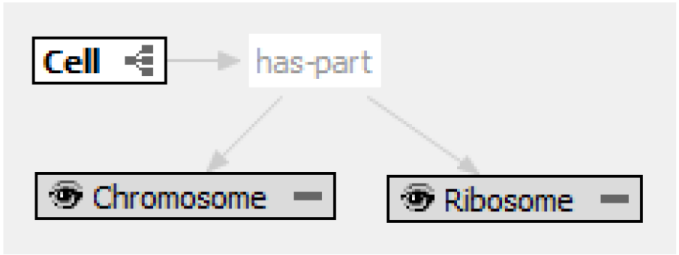
S2 Every *EukaryoticCell* is a *Cell*.

S3 Every *EukaryoticCell* has part a *EukaryoticRibosome*, a *EukaryoticChromosome*, a *Nucleus*, such that the *EukaryoticChromosome* is inside the *Nucleus*.

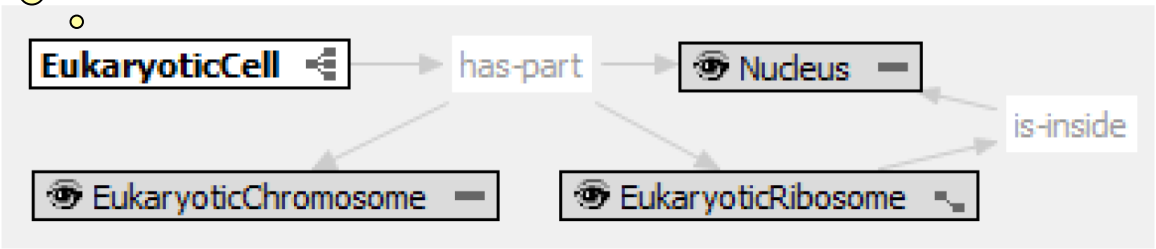
S4 Every *EukaryoticRibosome* is a *Ribosome*.

S5 Every *EukaryoticChromosome* is a *Chromosome*.

is-a edge implicit



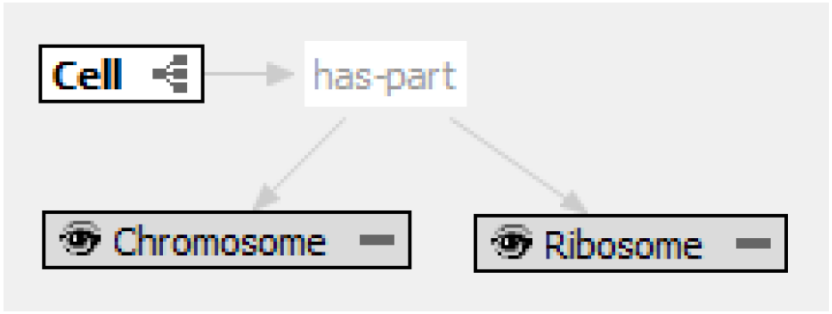
Same Ribosome that S1 is referring to? "Co-Reference Resolution"

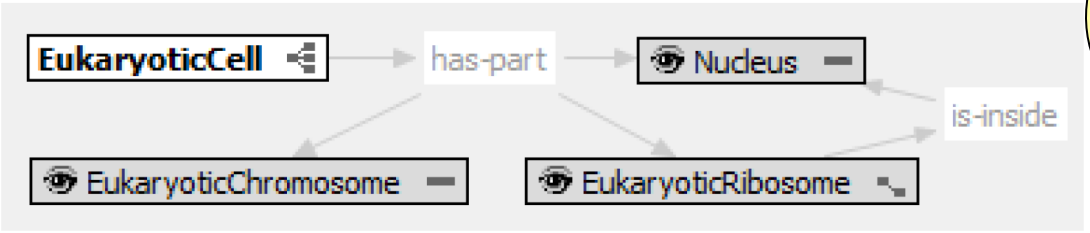


PROJECT HALO

“Underspecified” KBs

AURA





Q: Ambiguity - is that the Ribosome inherited from Cell super class?

A: Maybe – there are models in which this is the case, and models in which this is not the case.

=> “underspecified KB”

S1 $\forall x : Cell(x) \Rightarrow \exists x_1, x_2 : hasPart(x, x_1), Ribosome(x_1), hasPart(x, x_2), Chromosome(x_2)$

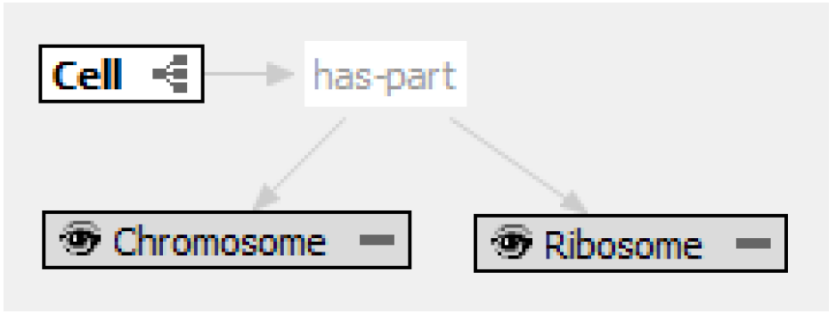
S23 $\forall x : EukaryoticCell(x) \Rightarrow \exists x_3, x_4, x_5 : Cell(x), hasPart(x, x_3), Euk.Ribosome(x_3), hasPart(x, x_4), Euk.Chromosome(x_4), hasPart(x, x_5), Nucleus(x_5), inside(x_4, x_5)$

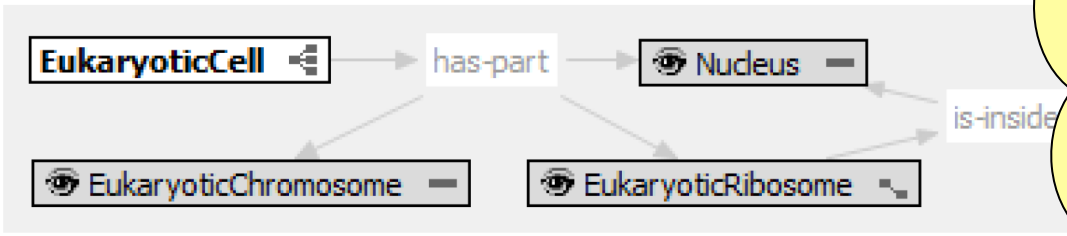
S4 $\forall x : Euk.Ribosome(x) \Rightarrow Ribosome(x)$

S5 $\forall x : Euk.Chromosome(x) \Rightarrow Chromosome(x)$

PROJECT HALO

Strengthened KBs





Q: Ambiguity - is that the Ribosome inherited from Cell super class?

A: yes! Due to "Skolem function inheritance" and equality, this holds in ALL models of the KB -> answer is entailed
=> "strengthened KB"

S1 ' $\forall x : Cell(x) \Rightarrow$
$hasPart(x, f_1(x)), Ribosome(f_1(x)),$
$hasPart(x, f_2(x)), Chromosome(f_2(x)),$

S23' $\forall x : EukaryoticCell(x) \Rightarrow Cell(x),$
$hasPart(x, f_3(x)), Euk.Ribosome(f_3(x)),$
$hasPart(x, f_4(x)), Euk.Chromosome(f_4(x)),$
$hasPart(x, f_5(x)), Nucleus(f_5(x)),$
$inside(f_4(x), f_5(x))),$
$f_3(x) = f_1(x), f_4(x) = f_2(x)$

Why do we care for strengthened KBs?

AURA

- **More entailments** (stronger KB / more deductive power)
- **Reduction of modeling effort** - suppose we extended Cell as follows:

In a Cell, every Ribosome is inside (a) Cytosol

S1b $\forall x : Cell(x) \Rightarrow \exists x_1, x_2, x_6 :$
$hasPart(x, x_1), Ribosome(x_1),$
$hasPart(x, x_2), Chromosome(x_2),$
$inside(x_1, x_6), Cytosol(x_6)$

underspecified

S1b' $\forall x : Cell(x) \Rightarrow$
$hasPart(x, f_1(x)), Ribosome(f_1(x)),$
$hasPart(x, f_2(x)), Chromosome(f_2(x)),$
$inside(f_1(x), f_0(x)), Cytosol(f_0(x))$

strengthened

only with **S1b'** can we deduce that this also holds for ***the*** EukaryoticRibosome in EukaryoticCell

- More entailed ("inherited") information – *hasPart(x, y1)* atom in **S23** is entailed from **{ S1b', S2 }**, but not from **{ S1b, S2 }**
- **Reduces KB size,** as entailed atoms are redundant
- **Provenance** ("from where is an atom inherited") is important for the modelers (Biologists in our case)

PROJECT HALO

This Work…

… presents an algorithm to construct a strengthened KB from an underspecified KB (**GSKB strengthening algorithm**)

S1 $\forall x : Cell(x) \Rightarrow \exists x_1, x_2 :$
$hasPart(x, x_1), Ribosome(x_1),$
$hasPart(x, x_2), Chromosome(x_2)$

S23 $\forall x : EukaryoticCell(x) \Rightarrow \exists x_3, x_4, x_5 : Cell(x),$
$hasPart(x, x_3), Euk.Ribosome(x_3),$
$hasPart(x, x_4), Euk.Chromosome(x_4),$
$hasPart(x, x_5), Nucleus(x_5), inside(x_4, x_5)$

S4 $\forall x : Euk.Ribosome(x) \Rightarrow Ribosome(x)$

S5 $\forall x : Euk.Chromosome(x) \Rightarrow Chromosome(x)$

S1b' $\forall x : Cell(x) \Rightarrow$
$hasPart(x, f_1(x)), Ribosome(f_1(x)),$
$hasPart(x, f_2(x)), Chromosome(f_2(x)),$
$inside(f_1(x), f_0(x)), Cytosol(f_0(x))$

S23' $\forall x : EukaryoticCell(x) \Rightarrow Cell(x),$
$hasPart(x, f_3(x)), Euk.Ribosome(f_3(x)),$
$hasPart(x, f_4(x)), Euk.Chromosome(f_4(x)),$
$hasPart(x, f_5(x)), Nucleus(f_5(x)),$
$inside(f_4(x), f_5(x))),$
$f_3(x) = f_1(x), f_4(x) = f_2(x)$

S4 $\forall x : Euk.Ribosome(x) \Rightarrow Ribosome(x)$

S5 $\forall x : Euk.Chromosome(x) \Rightarrow Chromosome(x)$

Note that this algorithm is not purely deductive by nature – it requires unsound reasoning namely hypothesization of equality atoms, NOT only Skolemization!

There may be more than one strengthened KB for a given underspecified KB.

Also note that the is-a relations and hence the taxonomy are given here. This is NOT a subsumption checking / classification problem!

Description Logics don't help for a variety of reasons (graph structures, unsound / hypothetical reasoning required, etc.)

The GSKB Strengthening Algorithm

AURA

Input: KB $\sum$ must be "admissible"
(no cycles -> finite model property)

Output: strengthened KB $\sum'$

1. **Skolemize KB $\sum$ -> KB $\sum_S$**
2. **Construct minimal Herbrand model of $\sum_S$: $(\Delta_{\mathcal{H}}, \cdot^{\mathcal{I}_{\mathcal{H}}})$**
3. **Use $(\Delta_{\mathcal{H}}, \cdot^{\mathcal{I}_{\mathcal{H}}})$ to construct a so-called preferred model of $\sum$: $(\Delta_{\mathcal{A}}, \cdot^{\mathcal{I}_{\mathcal{A}}})$**

   This step is non-deterministic, and it requires guessing of equalities. $\Delta_{\mathcal{A}} = \Delta_{\mathcal{H}} \backslash =$ is the quotient set of the Herbrand universe under those "guessed" equalities (=).
4. **Use $(\Delta_{\mathcal{A}}, \cdot^{\mathcal{I}_{\mathcal{A}}})$ and $\sum_S$ to construct $\sum'$**

PROJECT HALO

Preferred Models – Intuition

AURA

S1 $\forall x : Cell(x) \Rightarrow \exists x_1, x_2 :$
$hasPart(x, x_1), Ribosome(x_1),$
$hasPart(x, x_2), Chromosome(x_2)$

S23 $\forall x : EukaryoticCell(x) \Rightarrow \exists x_3, x_4, x_5 : Cell(x),$
$hasPart(x, x_3), Euk.Ribosome(x_3),$
$hasPart(x, x_4), Euk.Chromosome(x_4),$
$hasPart(x, x_5), Nucleus(x_5), inside(x_4, x_5)$

S4 $\forall x : Euk.Ribosome(x) \Rightarrow Ribosome(x)$

S5 $\forall x : Euk.Chromosome(x) \Rightarrow Chromosome(x)$

1. **In a preferred model, the concept models have the form of non-overlapping connected graphs, one node per variable**

2. **For every concept, there is at least one unique model which instantiates only this concept and its superconcepts, no other concepts - e.g., there is a model of Cell which is NOT also a model of EukaryoticCell**

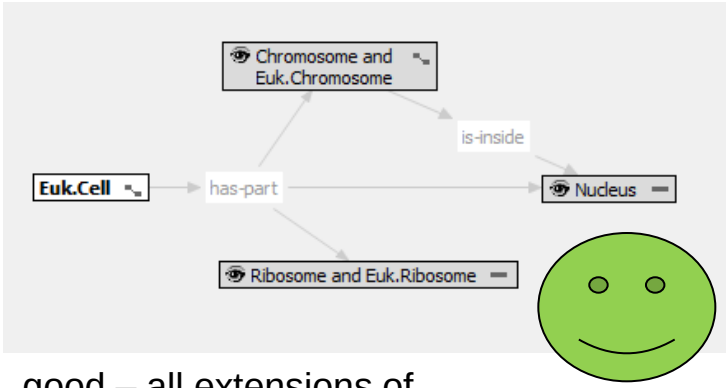
3. **In those concept models, the extensions of (possibly singleton) conjunctions are minimized – i.e., there is no admissible model which has a smaller extension for that conjunction. This forces us to identify successors "inherited from superclasses" with "locally specialized" versions**

PROJECT HALO

Models and Preferred Models

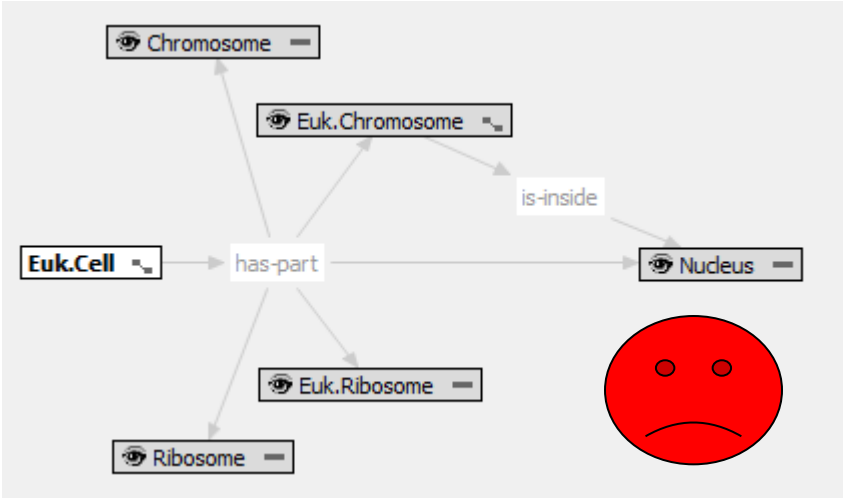
AURA

1



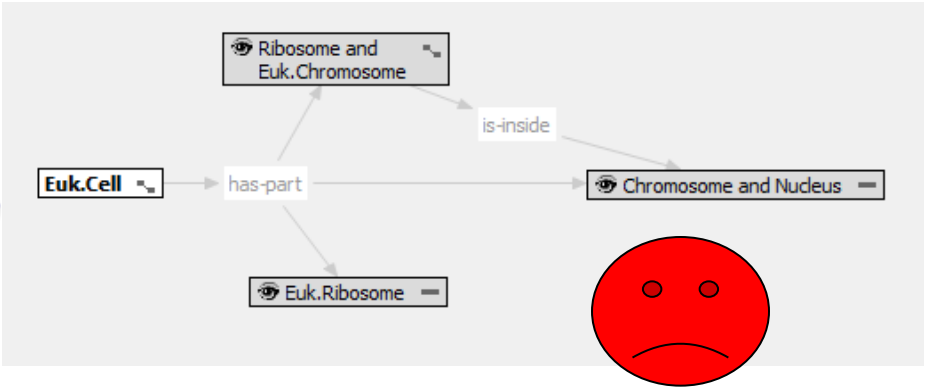
good – all extensions of (singleton) conjunctions minimal!
This is a preferred model !

2



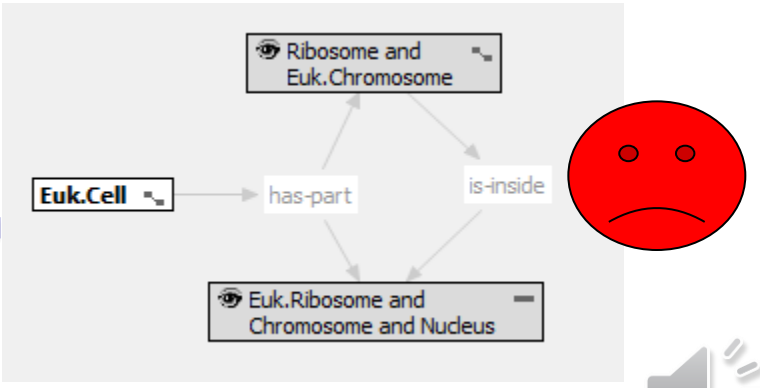
… too many Ribosomes and Chromsomes…

3



… non-empty extension of conjunction **Ribosome /\ Euk.Chromosome** (there are smaller models in which this conjunction is empty!)

4



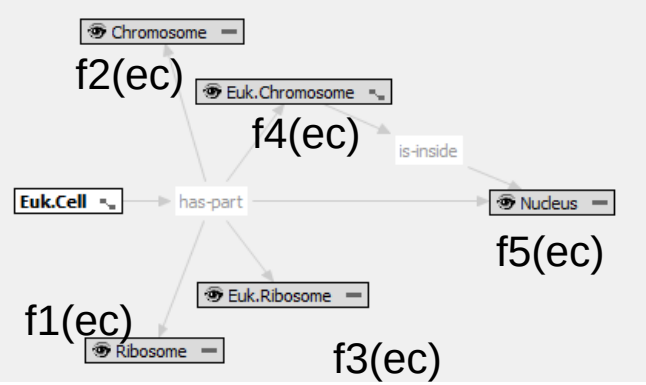
… even this is a model, but similar problems: non-empty conjunctions without necessity

PROJECT HALO

Constructing a Preferred Model

AURA

- **Start with the Herbrand model – this will satisfy conditions 1 and 2 of the admissible model**

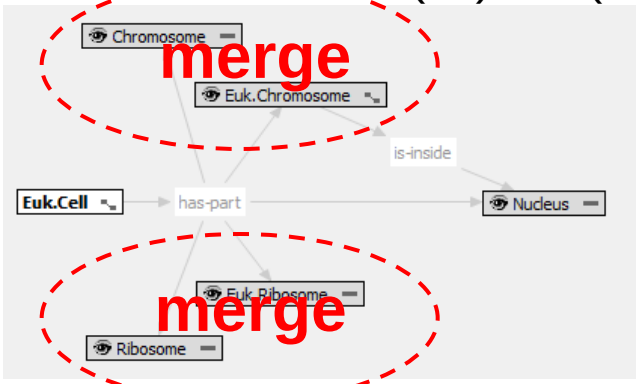


S1b' $\forall x : Cell(x) \Rightarrow$
$hasPart(x, f_1(x)), Ribosome(f_1(x)),$
$hasPart(x, f_2(x)), Chromosome(f_2(x)),$
$inside(f_1(x), f_0(x)), Cytosol(f_0(x))$

S23' $\forall x : EukaryoticCell(x) \Rightarrow Cell(x),$
$hasPart(x, f_3(x)), Euk.Ribosome(f_3(x)),$
$hasPart(x, f_4(x)), Euk.Chromosome(f_4(x)),$
$hasPart(x, f_5(x)), Nucleus(f_5(x)),$
$inside(f_4(x), f_5(x))),$

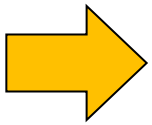
- **Identify and merge compatible successors using a non-deterministic merge rule, apply it exhaustively, and record in equality relation "="**

f2(ec) = f4(ec)

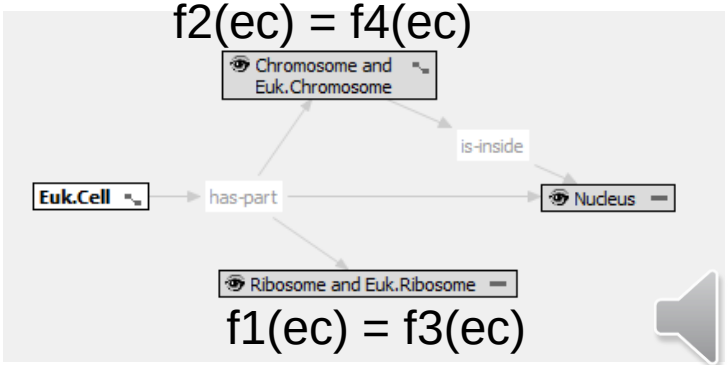


f1(ec) = f3(ec)

$\Delta_{\mathcal{H}}$



f2(ec) = f4(ec)



f1(ec) = f3(ec)

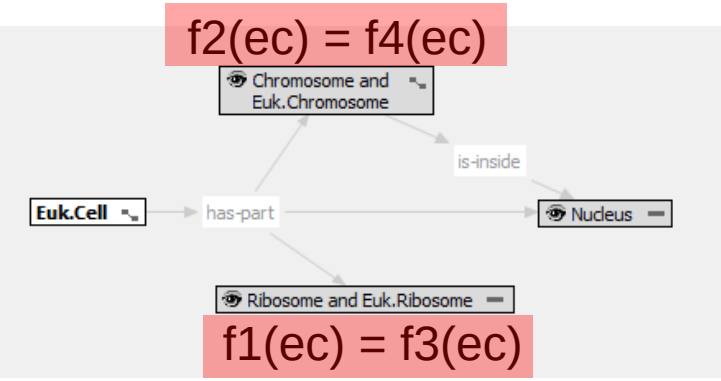
$\Delta_{\mathcal{A}} = \Delta_{\mathcal{H}} \setminus =$

PROJECT HALO

Constructing a Strengthened KB

- **For construction of the preferred model, the merge rule has been applied exhaustively**
  - **this has maximized the congruence / equality relation "="**

- **Now we simply add the equalities in "=" as equality atoms to the skolemized KB**



$$\Delta_{\mathcal{A}} = \Delta_{\mathcal{H}} \setminus =$$

S1b' $\forall x : Cell(x) \Rightarrow$
$hasPart(x, f_1(x)), Ribosome(f_1(x)),$
$hasPart(x, f_2(x)), Chromosome(f_2(x)),$
$inside(f_1(x), f_0(x)), Cytosol(f_0(x))$

S23' $\forall x : EukaryoticCell(x) \Rightarrow Cell(x),$
$hasPart(x, f_3(x)), Euk.Ribosome(f_3(x)),$
$hasPart(x, f_4(x)), Euk.Chromosome(f_4(x)),$
$hasPart(x, f_5(x)), Nucleus(f_5(x)),$
$inside(f_4(x), f_5(x))),$
$f_3(x) = f_1(x), f_4(x) = f_2(x)$

- $\sum'$ **is a strengthened KB and has preferred models**

Experiments

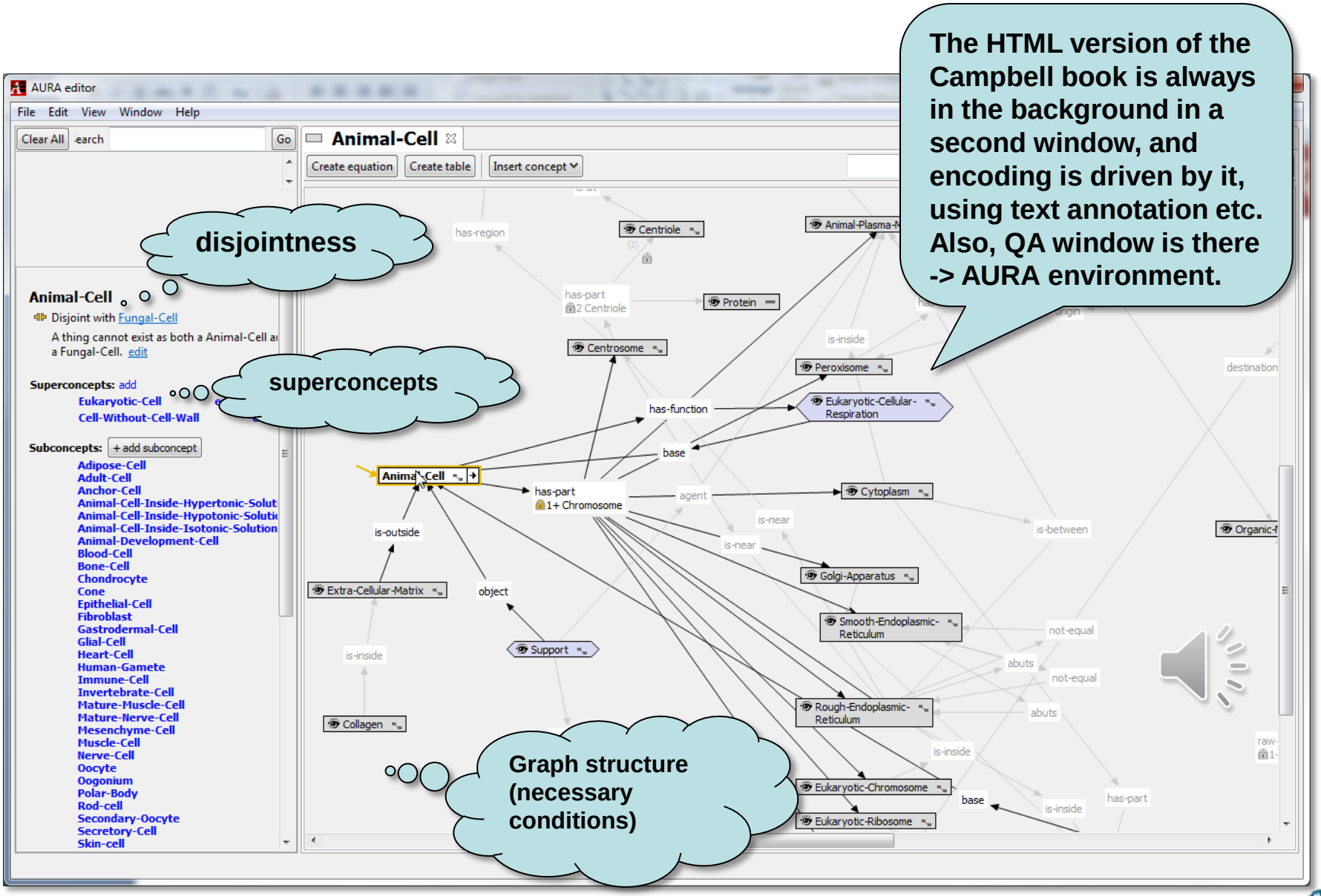
AURA

- We have a working KB strengthening algorithm which was applied to the AURA KB: it identified 82% of the 141,909 atoms as inherited and hypothesized 22,667 equality atoms. Runtime: 15 hours
- The algorithm works differently than described here, but the presented model-theoretic framework is a first step towards a logical formal reconstruction of the algorithm
- The native KR&R language of AURA is “Knowledge Machine” (KM)
  - the exploited KM representation does not support arbitrary equality atoms, hence this algorithm
- The actual implemented algorithm can handle additional expressive means, not yet addressed by the formal reconstruction (future work)
- The strengthened KB is also the basis for the AURA KB exports which are available for download!

PROJECT HALO

AURA Graphical Knowledge Editor

AURA



PROJECT HALO

AURA KB Stats (LATEST)

AURA

Regarding Class Axioms:

# Classes	# Relations	# Constants	Avg. # Skolems / Class	Avg. # Atoms / Necessary Condition	Avg. # Atoms / Sufficient Condition
6430	455	634	24	64	4

# Constant Typings	# Taxonomical Axioms	# Disjointness Axioms	# Equality Assertions	# Qualified Number Restrictions
714	6993	18616	108755	936

Regarding Relation Axioms:

# DRAs	# RRAs	# RHAs	# QRHAs	# IRAs	# 12NAs / # N21As	# TRANS + # GTRANS
449	447	13	39	212	10 / 132	431

Regarding Other Aspects:

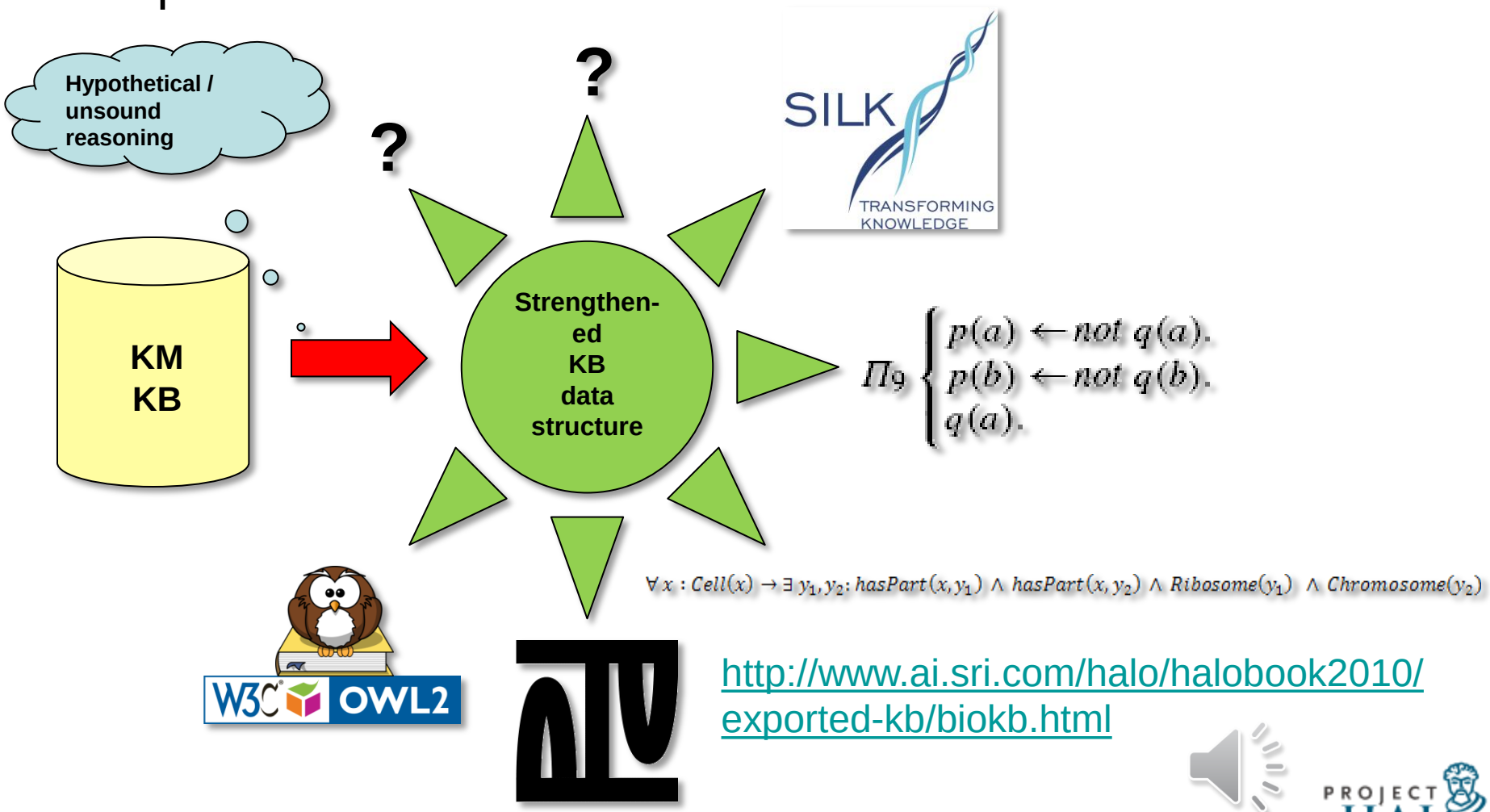
# Cyclical Classes	# Cycles	Avg. Cycle Length	# Skolem Functions
1008	8604	41	73815

PROJECT HALO

The Strengthened KB and AURA Exports

AURA

From the underlying KM representation, we are constructing the strengthened KB, which then gets exported into various standard formats



http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html

PROJECT HALO

Conclusion

AURA

- Strengthened GSKBs are important for a variety of reasons
  - to maximize entailed information / deductive power
  - to reduce KB size
  - to show correct provenance of atoms (inherited? local?) to KB authors
- Authoring strengthened KBs can be tedious or impossible (if the input is underspecified in the first place), hence an automatic strengthening algorithm is required
  - this is an unsound / hypothetical reasoning process which requires guessing of equalities
- We have presented first steps towards a formalization & logical reconstruction of an algorithm which solved an important application problem in the AURA project
  - our formalization is model-theoretic in nature and presents and exploits a novel class of preferred models
- As a by-product of these efforts, the AURA KB can now be exported into standard formats and KB_Bio_101 is available for download

PROJECT HALO

http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html

Thank you!

AURA Team in 2011