csc343 winter 2020
assignment #1: relational algebra
due February 7th, 4 p.m.

goals

This assignment aims to help you learn to:

- read a relational scheme and analyze instances of the schema
- read and apply integrity constraints
- express queries and integrity constraints of your own
- think about the limits of what can be expressed in relational algebra

Your assignment must be typed to produce a PDF document **a1.pdf** (hand-written submissions are not acceptable). You may work on the assignment in groups of 1 or 2, and submit a single assignment for the entire group on MarkUs. You must establish your group well before the due date by submitting an incomplete, or even empty, submission.

background

You will be working on a schema and queries for a database used by a zoological institute to track an archive of their artifacts.

During a field trip collectors gather a variety of artifacts of the animals they study, resulting in tissue samples, images, physical models (such as casts of paw prints), or live colonies.

After arriving at the institute, artifacts must be safely stored and maintained by technicians. Some artifacts are cited in one or more publications. In all cases the official species name must be recorded, and must appear in the Catalogue of Life database. If correct taxonomic practices are followed, each species belongs to exactly one genus, and each genus to exactly one family. Tables COL, Genus, and Species are derived from Catalogue of Life database.

relations

- Collection(CID, date, SID)
  Tuples here represent entire collections from a field trip, where *CID* is the collection ID, *date* is the starting date of the field trip, and *SID* is the staff ID of the collector.

- Collected(CID, AN)
  A tuple here represents the fact that collection *CID* includes artifact number *AN*. A single collection usually contains multiple artifacts, and a single artifact may be aggregated from more than one collection.

- Artifact(AN, species, type, location, SID)
  Tuples here represent single artifact collected in the field. *AN* is the artifact number, *species* is the scientific species name, *type* is one of tissue, image, model, or live, *location* is where it was collected, and *SID* is the staff number of the technician who maintains this artifact.

- Published(AN, journal, date)
  A tuple here represents the fact that artifact *AN* was mentioned in scholarly publication *journal* with publication date *date*.

- Staff(SID, name, email, rank, date)
  These tuples represent a member of the institute's scientific staff. *SID* is the staff ID, *name* is their full name, *email* is their professional email, *rank* is one of: technician, student, pre-tenure, or tenured, and *date* is the date when they attained that rank.

- COL(family)
  A singleton tuple here means that *family* is a scientific zoological family name that appears in the Catalogue of Life.

- Genus(genus, family)
  A tuple here means that *genus* is in family *family*.

- Species(species, genus)
  A tuple here means that *species* is in genus *genus*.

our constraints

For each of the following constraints give a one sentence explanation of what the constraint implies, and why it is required.

- $\pi_{species}(Artifact) - \pi_{species}(Species) = \emptyset$.
- $\pi_{rank}(Staff) \subseteq \{\text{'technician', 'student', 'pre-tenure', 'tenure'}\}$.
- $\pi_{family}(Genus) - \pi_{family}(COL) = \emptyset$.
- $\pi_{genus}(Species) \subseteq \pi_{genus}(Genus)$.
- $\pi_{CID}(Collected) = \pi_{CID}(Collection)$.
- $\pi_{AN}(Artifact) = \pi_{AN}(Collected)$.
- $\pi_{SID}(Collection) \subseteq \pi_{SID}(Staff)$.
- $\pi_{SID}(Artifact) \subseteq \pi_{SID}(Staff)$.
- $\pi_{type}(Artifact) \subseteq \{'tissue',' image',' model',' live'\}$
- $\pi_{AN}(Published) \subseteq \pi_{AN}(Artifact)$

queries

Write relational algebra expressions for each of the queries below. You must use notations from this course and operators:

$$\pi, \sigma, \rho, \bowtie, \bowtie_{condition}, \times, \cap, \cup, -, =$$

You may also use constants:

today (for current date) $\emptyset$ (for the empty set)

In your queries pay attention to the following:

- All relations are sets, and you may only use relational algebra operators covered in Chapter 2 of the course text.
- Do not make assumptions that are not enforced by our constraints above, so your queries should work correctly for any database that obeys our schema and constraints.
- Other than constants such as 23 or ”lupus”, a select operation only examines values contained in a tuple, not aggregated over an entire column.
- Your selection conditions can use arithmetic operators, such as $+, \leq, \neq, \geq, >, <$ and friends. You can use logical operators such as $\vee, \wedge$, and $\neg$, and treat dates and numeric attributes as numbers that you can perform arithmetic on.
- Use good variable names and provide lots of comments to explain your intentions.
- Return multiple tuples if that is appropriate for your query.

There may be a query or queries that cannot be expressed in the relational algebra you have been taught so far, in which case just write “cannot be expressed.” The queries below are not in any particular order.

1. Rationale: Performance reviews include seeing how current the work is of staff who have held their current rank for a long time.

   Query: Find the most recent collection date of any artifact collected by a staff member who has held their current rank the longest. Keep ties.

2. Rationale: Staff who maintain every artifact in some collection should be considered favourably in performance reviews.

   Query: Find all staff who maintain all artifacts in at least one collection.

3. Rationale: An artifact collected and maintained by the same staff may have some special requirements that should be investigated.

   Query: Find all artifacts that were collected by the same staff who maintains them.

4. Rationale: Identify multi-talented field workers.

   Query: Find all staff who have collected at least 3 artifacts from every species in some family.

5. Rationale: Which publications might have some specialized niche focus?

   Query: Find all publications that have used exactly 2 of our artifacts.

6. Rationale: Identify motherlode locations.

   Query: Find all locations where at least one artifact from every family has been collected.

7. Rationale: Exclusively tissue sample collectors may need extra support for special reagents and shipping costs.

   Query: Find all staff who have collected only tissue samples.

8. Rationale: Collection staff who should be encouraged to diversify their network.

   Query: Find all staff pairs who have worked only with each other on collections.

9. Rationale: Track the influence of a given staff member.

   Query: Staff member SID_1 is influenced by staff member SID_2 if (a) they have ever worked together on a collection or (b) if SID_1 has ever worked with a staff member who is influenced by SID_2. Find SIDs of staff members influenced by SID 42.

your constraints

For each of these constraints you should derive a relational algebra expression of the form $R = \emptyset$, where R may be derived in several steps, by assigning intermediate results to a variable. If the constraint cannot be expressed in the relational algebra you have been taught, write "cannot be expressed."

1. No species is also a genus.

2. No genus belongs to more than one family.

3. All publications must be published after all artifacts they use have been collected.

4. Students may not catalogue live artifacts.

submissions

Submit **a1.pdf** on MarkUs. One submission per group, whether a group is one or two people. You declare a group by submitting an empty, or partial, file, and this should be done well before the due date. You may always replace such a file with a better version, until the due date.

Double check that you have submitted the correct version of your file by downloading it from MarkUs.

marking

We mark your submission for correctness, but also for good form:

- For full marks you should add comments to describe the *data*, rather than *technique*, of your queries. These may help you get part marks if there is a flaw in your query.

- Please use the assignment operator, ":=" for intermediate results.

- Name relations and attributes in a manner that helps the reader remember their intended meaning.

- Format the algebraic expressions with line breaks and formatting that help make the meaning clear.