

Omar ElZaria

1004095739

Bingran Li

1004191910

assignment #1: relational algebra

Our Constraints

- This constraint implies that every artifact tuple must have a species field that matches a species field in Species.
This is required because you can't have an artifact that doesn't belong to a species.
- This constraint implies that every staff will have a rank field that matches one of the following: technician, student, pre-tenure, tenure.
This is required to describe what each staff member's role is.
- This constraint implies that for every genus tuple, there must exist a family field that matches a corresponding family field in COL.
This is required because you can't have a genus that doesn't belong to a scientific zoological family.
- This constraint implies that for every species tuple, the genus field must map to at least one genus from the Genus table.
This is required because you can't have a species that belongs to a genus that doesn't exist.
- This constraint implies that every artifact tuple contains an AN field matches to exactly one AN field in the collected table.
This is required because you can't have duplicate artifacts and there is only one AN for every artifact collected.
- This constraint implies that the number of staff IDs in the collection table can't exceed the number of staff IDs in the staff table.
This is required to ensure that every Staff associated with a collection is working for the zoological institute.
- This constraint implies that the number of staff IDs in the artifact table can't exceed the number of staff IDs in the staff table.
This is required to ensure that every Staff that's maintaining an artifact is working for the zoological institute.
- This constraint implies that every artifact will have a type that matches one of the following: tissue, image, model, live.
This is required to describe the type of different artifacts.

Omar ElZaria

1004095739

Bingran Li

1004191910

- This constraint implies that for every artifact that's mentioned in a scholarly publication, it maps to an artifact in the artifact table. This is required because you can't have an artifact that wasn't collected in the field to be mentioned in a scholarly publication.

Queries

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.

Solution:

First, we construct a table of all staff members excluding those who have held their current rank the longest:

$$\text{NotBottom}(\text{SID}) := \Pi_{s1, \text{SID}} (\sigma_{s1.date \geq s2.date} (\rho_{s1}(\text{Staff}) \times \rho_{s2}(\text{Staff})))$$

Second, we subtract that table from the staff table to end up with a table which holds all staff who has held their current rank the longest:

$$\text{OldestStaff}(\text{SID}) := \Pi_{\text{SID}} (\text{Staff} - \text{NotBottom})$$

Third, we find the most recent collection date of any artifact collected by the staff member using the "Find Max" method:

$$\text{ArtifactsCollected}(\text{SID}, \text{date}) := \Pi_{\text{SID}} (\text{OldestStaff} \bowtie \text{Collection})$$

$$\begin{aligned} \text{NotRecentArtifact}(\text{SID}, \text{date}) := & \Pi_{A2, \text{SID}} (\sigma_{A1.date \geq A2.date} (\rho_{A1}(\text{ArtifactsCollected}) \times \rho_{A2} \\ & (\text{ArtifactsCollected}))) \end{aligned}$$

$$\text{RecentArtifact}(\text{SID}, \text{date}) := (\Pi_{\text{SID}, \text{date}} \text{ArtifactsCollected}) - \text{NotRecentArtifact}$$

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.

Solution:

First, we natural join the Collected and Artifact tables to project a table that holds collections IDs and the IDs of the staff maintaining those collections:

$$\text{CollectionStaff}(\text{CID}, \text{SID}) := \Pi_{\text{CID}, \text{SID}} (\text{Collected} \bowtie \text{Artifact})$$

Second, we remove any duplicates to obtain a table of distinct staff IDs and project the names of all staff who maintain all artifacts in at least one collection:

Omar ElZaria

1004095739

Bingran Li

1004191910

$$\text{Duplicates}(\text{CID}, \text{SID}) := \Pi_{a.\text{CID}, a.\text{SID}} (\sigma_{a.\text{CID}=b.\text{CID}, a.\text{SID} \neq b.\text{SID}} (\rho_a (\text{CollectionStaff}) \times \rho_b (\text{CollectionStaff})))$$

$$\text{SingleCollectionStaff}(\text{SID}) := \Pi_{\text{SID}} (\text{CollectionStaff} - \text{Duplicates})$$

$$\Pi_{\text{name}} (\text{SingleCollection} \bowtie \text{Staff})$$

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.

Solution:

First, we natural join the Collected and Collection tables to obtain a table holding all the artifact numbers and Staff IDs. Second, we use a theta join with the Artifact table on the condition that the artifact number match as well as the staff ID of the collector matches the Staff ID of the staff member maintaining that artifact:

$$\Pi_{a.\text{AN}} (\rho_a (\Pi_{\text{AN}, \text{SID}} \text{Artifact}) \bowtie_{a.\text{AN}=b.\text{AN}, a.\text{SID}=b.\text{SID}} \rho_b (\Pi_{\text{AN}, \text{SID}} \text{Collected} \bowtie \text{Collection}))$$

4. **Rationale:** Identify multi-talented field workers.

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

Solution:

First, we create a table that holds AN, species and family through the following sequence of natural joins:

$$\text{RequiredData}(\text{AN}, \text{species}, \text{family}) := (\Pi_{\text{AN}, \text{species}, \text{family}} (\text{Artifact} \bowtie (\text{Species} \bowtie \text{Genus}) \bowtie (\Pi_{\text{AN}, \text{SID}} \text{Collection} \bowtie \text{Collected})))$$

Then, we select the tuples of all staff who have collected at least 3 artifacts from every species in some family:

$$\sigma_{a.\text{species} = b.\text{species}, a.\text{species} = c.\text{species}, b.\text{species} = c.\text{species}, a.\text{AN} \neq b.\text{AN}, a.\text{AN} \neq c.\text{AN}, b.\text{AN} \neq c.\text{AN}, a.\text{SID} = b.\text{SID}, a.\text{SID} = c.\text{SID}, b.\text{SID} = c.\text{SID}} (\rho_a (\text{RequiredData}) \times \rho_b (\text{RequiredData}) \times \rho_c (\text{RequiredData}))$$

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

Omar ElZaria

1004095739

Bingran Li

1004191910

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

Solution:

First, we created table that holds pairs of artifacts that were mentioned in the same journal to help us narrow down to the required query:

$$\text{ArtifactPairs}(\text{AN1}, \text{AN2}, \text{journal}) := \Pi_{\text{AN1}, \text{AN2}, \text{a.journal}} (\sigma_{\text{AN1} > \text{AN2}, \text{a.journal} = \text{b.journal}} (\rho_{\text{a}(\text{AN1}, \text{journal}, \text{date})} (\text{Published}) \times \rho_{\text{b}(\text{AN2}, \text{journal}, \text{date})} (\text{Published})))$$

Second, we found all journals that had 2 or more artifacts mentioned in them:

$$\text{TwoOrMore}(\text{journal}) := \Pi_{\text{journal}} (\text{Artifact}) \bowtie \Pi_{\text{journal}} (\text{Artifact})$$

Third, we found all journals that had more than 2 artifacts mentioned in them:

$$\text{MoreThanTwo}(\text{journal}) := \Pi_{\text{a.journal}} (\sigma_{\text{a.AN1} = \text{b.AN1}, \text{a.AN2} \neq \text{b.AN2}, \text{a.journal} = \text{b.journal}} (\rho_{\text{a}} (\text{ArtifactPairs}) \times \rho_{\text{b}} (\text{ArtifactPairs})))$$

Finally, we found all publications that have used exactly 2 artifacts by subtracting the previous two tables that we have created:

$$\text{ExactlyTwo}(\text{journal}) := \text{TwoOrMore} - (\text{MoreThanTwo} \bowtie \text{MoreThanTwo})$$

6. Rationale: Identify motherlode locations.

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

Solution:

First, we natural joined Artifact, Species and Genus tables to obtain a table that holds location and family. We also eliminated any duplicates in that table:

$$\text{LocData}(\text{location}, \text{family}) := \Pi_{\text{location}, \text{family}} ((\text{Artifact} \bowtie \text{Species}) \bowtie \text{Genus})$$

$$\text{LocDataNoDup}(\text{location}, \text{family}) := \Pi_{\text{location}, \text{family}} (\text{LocData} \bowtie \text{LocData})$$

$$\text{LocOnly}(\text{location}) := \Pi_{\text{location}} (\text{LocData}) \bowtie \Pi_{\text{location}} (\text{LocData})$$

Second, we created a table that holds all the locations where at least one artifact from every family was not found:

$$\text{AtLeastOneNotFound}(\text{location}, \text{family}) := (\text{LocOnly} \times \text{COL}) - \text{LocDataNoDup}$$

Omar ElZaria
1004095739
Bingran Li
1004191910

$$\text{LocOnlyNotFound}(\text{location}) := \Pi_{\text{location}} (\text{AtLeastNotFound}) \bowtie \Pi_{\text{location}} (\text{AtLeastNotFound})$$

Finally, we subtracted these two tables in order to end up with a table that holds all motherlode locations

$$\text{Motherlodes}(\text{location}) := \text{LocOnly} - \text{LocOnlyNotFound}$$

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.

Solution:

First, we create a table that holds staff IDs and the corresponding type of artifact that each staff has collected:

$$\text{CollectorData}(\text{SID}, \text{Type}) := \Pi_{\text{SID}, \text{type}} (\text{Collection} \bowtie \text{Collected}) \bowtie \Pi_{\text{AN}, \text{type}} (\text{Artifact})$$

Second, we create a table that holds the SIDs of all staff that haven't collected any artifacts of type tissue:

$$\text{CollectNoTissue}(\text{SID}) := \Pi_{\text{SID}} (\sigma_{\text{type} \neq \text{'tissue'}} (\text{CollectorData}))$$

Finally, we create a table that holds the SIDs of all staff who have collected only tissue samples

$$\begin{aligned} \text{OnlyTissue}(\text{SID}) := & (\Pi_{\text{SID}} (\text{CollectorData}) \bowtie \Pi_{\text{SID}} (\text{CollectorData})) - (\text{CollectorNoTissue} \\ & \bowtie \text{CollectorNoTissue}) \end{aligned}$$

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.

Solution:

First, we create a table that holds pairs of staff members by natural joining the Artifacts and Collected tables twice to make two new separate tables and then crossing these two tables together to project pairs of SIDs:

$$\begin{aligned} \text{StaffPairs}(\text{staff1}, \text{staff2}) := & \Pi_{a.\text{SID}, b.\text{SID}} (\sigma_{a.\text{CID} = b.\text{CID}} (\rho_a (\text{Artifacts} \bowtie \text{Collected}) \times \rho_b \\ & (\text{Artifacts} \bowtie \text{Collected}))) \end{aligned}$$

Omar ElZaria

1004095739

Bingran Li

1004191910

After that, we find all the duplicate staff pairs contained in the table above and subtract those duplicates from the table above to project a table that holds all staff pairs who have worked only with each other on collections:

$$\text{DupStaff}(\text{staff1}, \text{staff2}) := \Pi_{a.\text{staff1}, a.\text{staff2}} (\sigma_{a.\text{staff1} = b.\text{staff1}, a.\text{staff2} \neq b.\text{staff2}} (\rho_a (\text{StaffPairs}) \bowtie \rho_b (\text{StaffPairs})))$$

$$\text{SmallNetwork}(\text{staff1}, \text{staff2}) := \text{StaffPairs} - \text{DupStaff}$$

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

Query: Staff member SID1 is influenced by staff member SID2 if (a) they have ever worked together on a collection or (b) if SID1 has ever worked with a staff member who is influenced by SID2. Find SIDs of staff members influenced by SID 42.

Solution:

Cannot be expressed

Your Constraints

1. No species is also a genus.

Solution:

$$\Pi_{\text{species}} (\text{Species}) \cup \Pi_{\text{genus}} (\text{Genus}) = 0$$

2. No genus belongs to more than one family.

Solution:

$$\text{FamiliesOnly} := \Pi_{\text{family}} (\text{Genus}) \bowtie \text{COL}$$

$$\text{FamiliesOnly} - \Pi_{\text{family}} (\text{COL}) = 0$$

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

Solution:

$$\text{ReqData} := \Pi_{\text{date1}, \text{date2}} ((\rho_{\text{Collection}(\text{CID}, \text{date1}, \text{SID})} (\text{Collection}) \bowtie \text{Collected}) \bowtie \Pi_{\text{AN}} (\text{Artifact})) \bowtie$$

$$\rho_{\text{Published}(\text{AN}, \text{journal}, \text{date2})} (\text{Published}))$$

$$\sigma_{\text{date2} < \text{date1}} (\text{ReqData}) = 0$$

Omar ElZaria

1004095739

Bingran Li

1004191910

4. Students may not catalogue live artifacts.

Solution:

$$\sigma_{type = 'live'}(\Pi_{type, SID} (Artifact) \bowtie \Pi_{SID} (\sigma_{rank = 'student'} Staff)) = 0$$