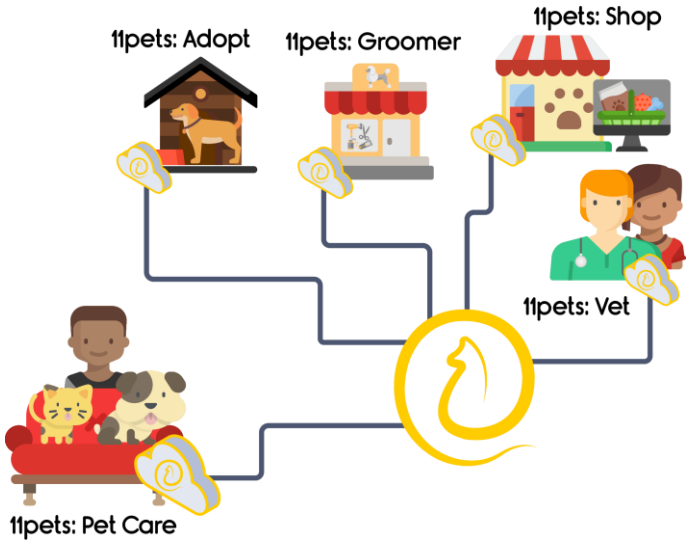
Data Management for Knowledge Extraction

Demos Pavlou (demos.pavlou@11pets.com), Georgios Moullotos (giorgos.moullotos@11pets.com), Kyriakos Stavrou (kyriakos.stavrou@11pets.com)

The requirements

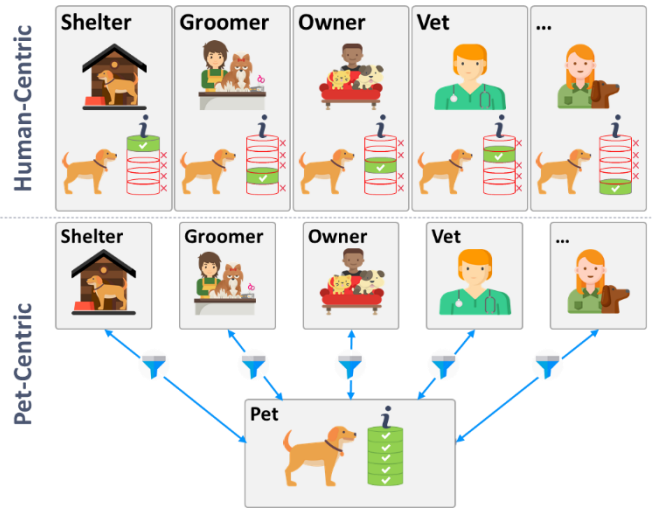
11pets is the leading software ecosystem for the pet industry offering solutions for pet families, pet professionals and pet welfare organizations. 11pets offers web-based and mobile application solutions that help the different actors manage the data of their pets and their day-to-day business needs.



Our large userbase and the wide usage of our system, allowed us to collect a significant volume of interdisciplinary data (>10M data points) from different sectors of pet-care including care at home, by veterinarians, groomers, and shelters. By analyzing this pool of data, we have published multiple industry reports that drew the attention of key enterprises. These reports, however, are only based on descriptive analysis of the data, that is, statistical analysis with no knowledge extraction.

The data fragmentation challenge

Today's systems are *human-centric*; they are built around the pet professionals rather than the pets and their families. The current approach is equivalent to a national health system where each doctor, hospital, school, etc. has its own, *isolated* copy of each person's data and is not able to share it with anyone else. This data fragmentation makes it impossible to provide optimal care as each entity has just a subset of the information and nobody can see the whole picture.



In a *pet-centric* system, there is one, consistent copy of the pet's data that includes all information with each entity having a different view. This allows structured and standardized sharing of information. Moving the center of pet-care from the service providers to the pet and its family allows collaboration between each entity. Teamwork means more information for professionals and efficiency in the identification and solution of issues providing the

breakthrough the industry needs. Imagine a pet that had an allergic reaction while in the shelter and its new vet is not aware. It is only in a pet-centric system that continuity is guaranteed and professionals have access to all the necessary information.

11pets was the first to introduce the pet-centric approach to the industry and is today the most widely used digital pet-care platform. The 11pets platform enables collaboration at every level considering all privacy constraints. Professionals and families select which data they will share, with whom, and for how long. Each time we release a new product and cover another sector of the industry, we enrich our data not only with additional information but also, with additional parameters.

Knowledge extraction using pet-care data

Machine learning and knowledge/intelligence extraction are widely used today in a variety of areas, including healthcare, security, transportation, risk detection, risk management, etc. [ref]. The field had major contributions in the areas of self-driving cars, natural language translation, and healthcare [ref]. It is a field that is growing rapidly and each day is applied to more real-life problems.

Recently, there have been several efforts on using data analysis for the pet industry [ref, ref]. However, these studies cover only one pet-care sector at a time as their datasets are limited to it. As explained already, modern pet-care involves different sectors that collaboratively care for each pet. The pet-centric nature of the 11pets platform allows all these care providers to work together by centralizing the data management. This, in turn, enables having composite, consistent and coherent interdisciplinary data from different sectors, making it unique. An example of a study that is only enabled by this pet-centric approach is the identification of the relationship between the care a pet receives at the groomer and any skin allergies later treated by a veterinarian.

The data layout of 11pets

For knowledge and intelligence extraction it is necessary to guarantee the consistency of the data, that is, all references to each entity should be unique in the whole system. Simply stated, all actors should see the same view each entity respectless of their role. For example, a veterinarian, groomer, of pet shop, should see the same data of the pet; when one actor modifies the data of a pet, all other actors should see the updated information.

There are three key conflicting requirements:

1. **Data consistency:** The different actors should see the same data at all times.
2. **Private data:** Each actor should be able to keep private data for each entity. For example, a groomer might want to keep additional information about a pet regarding its grooming particularities. This dataset is private to the groomer and should neither be visible, nor modifiable by the other actors.
3. **Personal data:** Each data point should have a well-defined owner. Legally, the data of a pet belongs to the family. However, private data (such as what was explained the previous point) should belong to the actor that added it.

The case of shared data

Consider the case that a pet family takes the pet to a veterinarian for an x-ray. The veterinarian takes the x-ray and adds it to the pet's data. The veterinarian analyzes the x-ray and writes the diagnosis and also, keep information about how he/she inferred the diagnosis from the x-rays. At the same time, the family adds a comment about the treatment it received from the veterinary clinic.

This scenario is very common and presents a complicated data-ownership scheme.

- **Appointment data:** The ownership belongs to the veterinarian and can be shared with the family.
- **X-ray:** The family has paid for the x-ray itself and has the right to obtain a copy of the file. The veterinary clinic, also has access to the file. The system should guarantee that even if the one party (e.g., the family) deletes or modifies the x-ray, the copy of the other party (the veterinarian in this example) will retain the original information.
- **Diagnosis:** This has exactly the same treatment as the x-ray.
- **Inference / internal veterinary clinic notes:** This piece of information is private to the clinic and forms part of the intellectual properly the system needs to preserve. The family should not have access to this information.
- **Family notes:** Similarly, any notes the family adds for the treatment, is private to it and should not be visible to the clinic.

The data management scheme

The data management scheme of 11pets is based on the following principles:

1. The baseline database schema allows one entry per entity (e.g., each pet is written only once)
2. The system maintains separate tables for the actor-specific data. This mechanism regards for example the specific information professionals add for the pet (e.g., grooming information for groomers, veterinary data for clinics etc.)
3. The different actors have separate views depending on their role. This is supported by a proprietary SLA (Service Layer Agreement) mechanism that defines the elements each actor can access and modify.

The sharing of data between the different actors is as follows:

1. The source actor shares the information with the destination actor (e.g., if a veterinarian will share an x-ray with the family, then the source actor is the veterinarian and the destination actor is the family).
2. When the sharing is submitted, the destination is given read-access to the data.
3. Updating of the information:
    a. **Source Update**: If the source updates the information, the destination views the updated version of the data
    b. **Destination Update**: If the destination wants to update the data, then a separate “notes-like” entry is created that is stored together with the original information. The source cannot access this part.
    c. **Deletion**: If the source or destination deletes the data, then they lose the permission to access the data. The other actor maintains the access permission.

Data sanitization

Human entered data offer suffer from mistakes which significantly limit the expressive power of the information. Some examples include:

- **Explicitly invalid information:**
    - For example, information about estrous cycles / pregnancies for neutered pets.
    - Prescription of medications for the wrong species

- **Outlier data:** Datapoints that are way different than other similar metrics (e.g., weight).

Explicitly Invalid information

The system periodically runs correlation analysis on the dataset and lists “suspicious cases”, that is, data correlations that were not seen before. 11pets is equipped with a special support mechanism that allows us to define these invalid conditions dynamically. The mobile apps periodically download these rules and warn the user when a potential error is detected.

Outlier Information

For all numerical values, the system performs and outlier analysis to identify the expected range. Notice that such analysis depends on the characteristic of the pet. As an example, the weight depends on the gender, age, species and breed.

Data sanitization for Machine learning processing

Notice that all these suspicious data points are marked by the system when they are seen for the first time and are never used to train our machine learning models or for other data analysis activities.

Data collection

Currently, we are collecting data for the execution overhead and efficiency of all these mechanisms. We expect to submit a relevant paper in the near future.

Acknowledgements

“This work was co-funded by the European Union and the Republic of Cyprus through the Research and Innovation Foundation (Project: INNOVATE-COVID/0420/0018)”.